colecția BIBLIOTECA BĂNCII NAȚIONALE

13

Nicholas
GEORGESCU-ROEGEN

Metoda statistică

III*

EDITURA Expert

Scanned with OKEN Scanner

Cartea de față, deși are la bază cursul predat studenților Școlii de Statistică și se adresează, prin urmare, în special acestora, cuprinde un câmp de probleme mai vast decât cel încadrat în lecțiile de școală.

Am căutat ca prin aceasta să pun la dispoziția celor pe care metodele moderne ale statisticii matematice i-ar putea interesa o lucrare românească utilă și pe cât posibil mai completă.

Planul după care a fost întocmită această lucrare a avut un dublu scop.

Să dea, în primul rând, celor cu o pregătire matematică elementară posibilitatea utilizării celor mai noi procedee de cercetare; în al doilea, să pună pe cercetătorul teoretic în contact intim cu aplicațiile practice ale statisticii matematice, îndemnându-l astfel să-și aleagă problemele dintre acelea care într-adevăr pot prezenta o utilitate practică. În acest din urmă scop, am adăugat la sfârșitul lucrării un apendice care nu se adresează decât celor cu o pregătire matematică înaintată. Astfel se și explică faptul că amănuntele demonstrațiilor au fost uneori lăsate de o parte. Lectura apendicelui nu este câtuși de puțin necesară pentru o înțelegere completă a primelor patru părți; este absolut necesar ca cei ce o vor întreprinde să fi citit textul acestora din urmă.

N. Georgescu-Roegen (1933)

Scanned with OKEN Scanner

COORDONATORUL COLECȚIEI „BIBLIOTECA BĂNCII NAȚIONALE”

Prof.univ.dr. Mugur Isărescu

EDITURA Expert

București, România

Toate drepturile asupra acestei ediții, ca și a întregii opere roegeniene în limba română (vol. I-VII) din prezenta colecție, aparțin Institutului Național de Cercetări Economice al Academiei Române. Reproducerea, fie și parțială și pe orice suport, este interzisă fără acordul scris al Institutului, fiind supusă prevederilor legii drepturilor de autor.

Scanned with OKEN Scanner

Nicholas GEORGESCU-ROEGEN

OPERE COMPLETE

III

Cartea I

METODA STATISTICĂ

ELEMENTE DE STATISTICĂ MATEMATICĂ

Ediția a II-a

Coordonare generală
Aurel IANCU
Membru corespondent al Academiei Române

EDITURA Expert

Scanned with OKEN Scanner

Nicholas GEORGESCU-ROEGEN
Statistical Method. Elements of Mathematical Statistics

Nicholas GEORGESCU-ROEGEN
Méthode statistique. Eléments de statistique mathématique

Nota redacţiei

Prima ediţie a acestei lucrări, elaborată de N. GEORGESCU-ROEGEN, profesor la Şcoala de Statistică, secretarul Institutului de Statistică Generală a Statului, a apărut sub egida Institutului, în anul 1933, cu o prefaţă a lui Octav Onicescu, profesor la Facultatea de Ştiinţe din Bucureşti, directorul Şcolii de Statistică.
Iconografia prezentei ediţii este reprodusă, prin procesare electronică, din ediţia iniţială.

Coperta colecţiei: *arhitect Alexandru Chialda*
Prezentare grafică: *Nicolae Login*

ISBN 973-9282-37-7
973-9282-38-5

Apărut 1998

Scanned with OKEN Scanner

Pentru devotamentul ei neșovăit și clar, pentru renunțarea și sacrificiile îndelung voite, această carte, semn și dreaptă închinare

Mamei mele

Scanned with OKEN Scanner

Cuprins

PREFAȚĂ, *Octav Onicescu* 9
O LĂMURIRE, *Autorul* 11

PARTEA I

CAPITOLUL I – Introducere 15
CAPITOLUL II – Variabile, unități, serii statistice 23
CAPITOLUL III – Reprezentarea grafică 49

PARTEA a II-a

CAPITOLUL IV – Valori tipice 91
CAPITOLUL V – Dispersie, asimetrie 129
CAPITOLUL VI – Distribuția normală, ajustare 155
CAPITOLUL VII – Legile hazardului 179
CAPITOLUL VIII – Analiza seriilor homograde – *Schemele lui Bernoulli, Poisson, Lexis* 203
CAPITOLUL IX – Comparația statistică 217

PARTEA a III-a

CAPITOLUL X – Asociație, contingență 239
CAPITOLUL XI – Regresie, corelație 255
CAPITOLUL XII – Alte aspecte ale corelației; corelație parțială 297

PARTEA a IV-a

CAPITOLUL XIII – Analiza seriilor în timp 327
CAPITOLUL XIV – Covariația 377
CAPITOLUL XV – Numere indice 399
APENDICE 415
TABLE NUMERICE 461
INDEX ALFABETIC 465
INDEX ONOMASTIC 469

Scanned with OKEN Scanner

Prefață

Într-un anume sens, se poate spune că sunt atâtea statistici câți statisticieni sunt.

Lucrul merge așa de departe în domeniile căzute mai de curând sub imperiul statisticii, cum e, de pildă, domeniul cercetărilor asupra constituției materiei, încât se vorbește acolo despre o statistică a lui Einstein, despre una a lui Bose, despre statistica lui Fermi, lăsând la o parte mai vechile și glorioasele statistici ale lui Maxwell și Boltzmann.

În mod fatal, statistica urmează condițiile speciale ale materialului obiectiv la al cărui studiu omul de știință o cheamă în ajutor. Și chiar dacă este vorba de un anume material obiectiv, ipoteze felurite asupra lui atrag un tratament metodic diferit, adesea chiar o creație ad-hoc de forme și de tipuri de calcul.

Această constatare este legată de o alta: cunoașterea realității nu este încă îngăduită decât în frânturi ale ei, cu mare osteneală izolate de noi pentru a le preface în obiect de cercetare.

Până în preajma acestui secol știința a trăit iluzia mecanicistă. Orice schemă deterministă, în ultimă analiză cauzală, ținea să-și găsească o interpretare mecanică. Și idealul științei, care nu părea nici absurd, nici depărtat, era această mecanică universală. Cu atomi (am zice acum electroni), cu iuțeli, cu accelerații și cu forțe urma să se construiască imaginea completă a universului. Greutățile care stăteau în cale trebuia să constea doar în rezolvarea unor ecuații diferențiale.

Astăzi, idealul acesta apare cu mult prea simplist. Am renunțat să construim cunoașterea lumii pe baza determinismului cauzal, cu atât mai mult să o considerăm ca un mecanism uriaș. Dar n-am renunțat la idealul unei teorii a cunoașterii totale, deși supunem acest ideal exigențelor unei teorii a cunoașterii statistice.

Este principial posibilă o atare cunoaștere? Are, adică, metoda de cercetare statistică o justificare obiectivă, nu în aspecte izolate și variate ale unor frânturi de experiență, ci în universul ca experiență unică, totală?

Am încercat, într-un studiu publicat de „Arhiva", să dăm un răspuns afirmativ acestor chestiuni fundamentale, arătând treptele cunoașterii statistice și condițiile ei.

Sub hainele variate ale feluritelor aspecte ale statisticii sunt un număr de elemente comune, care constituie schema științei statistice. Metodă, în primul rând, adaptată însă în așa măsură universului, încât ambiționează să-i pătrundă cele mai ascunse și mai complexe aspecte și să realizeze cunoașterea lui integrală. Cu valori tipice, cu frecvențe, cu fluctuații, cu corelații și cu indici, cu câteva scheme tipice, statistica speră să dea imaginea lumii, făcând să renască,

Scanned with OKEN Scanner

pentru omul de ştiinţă, iluziile pe care le-a trăit mecanicismul veacului trecut. Cu mai multă înţelepciune însă, făcând loc şi acelor necesare limite ale cunoaşterii ştiinţifice în mic, spre lumea microscopică, în care principiul nedeterminării lui Heisenberg sau însăşi teoria cuantelor stabilesc marginile posibile ale măsurătorilor, în mare, spre lumea numerelor mari, care constituie prin ea însăşi obiect nou de cunoştinţă diferit de elementele constitutive şi aproape esenţial străin acestora, deci eventual însăşi comprehensiunii lor.

Pornind de la aceste postulate, lucrarea de faţă prezintă materialul statisticii într-o înlănţuire sistematică şi organic gândită.

Ea cuprinde noţiunile statistice fundamentale, metodele statistice cele mai obişnuite, atât relativ la descrierea fenomenelor statistice şi analiza lor, cât şi la cercetarea relaţiilor statistice dintre ele, cu referinţă specială la unele fenomene economice şi la acelea care constituie biometrica, fenomene care formează nu numai un obiect de preferinţă a autorului, dar şi un model al oricărei statistici.

Cartea de faţă reprezintă în esenţă cursul de statistică matematică pe care autorul îl face, din însărcinarea mea, la Şcoala de Statistică.

Studiile teoretice, aprofundate şi variate în acelaşi timp, cu care dl N.Georgescu-Roegen, statistician de la Institutul de Statistică al Universităţii din Paris, a trecut prin University College de la Londra, lucrând cu Pearson, l-au indicat de la sine pentru o atare însărcinare, ca şi pentru aceea reprezentată de această lucrare. Cu atât mai mult cu cât teoreticianului i se asociază, prin funcţia pe care o are la Institutul de Statistică, statisticianul practic, cunoscând valoarea teoriilor, a formulelor, în măsură să discearnă teoriile utile de simple jocuri de formule sau de teoretizare formală, deci în măsură să facă acea selecţie şi valorificare care singură dă preţ unei cărţi de ansamblu.

Autorul a lăsat pentru apendice justificarea matematică a câtorva noţiuni, legături de noţiuni sau ale unor afirmaţii făcute în desfăşurarea capitolelor cărţii, pentru a nu strica simplitatea planului şi a nu abate prin aceasta atenţia asupra unor consideraţii care interesează mai ales pe matematicieni.

Şcoala de Statistică nu era în măsură să suporte cheltuielile tiparului. Institutul de Statistică Generală a Statului a binevoit să pună la dispoziţie o parte din puţinele fonduri destinate tipăriturilor sale, cu înţelegătoarea aprobare a autorităţilor superioare cărora şi în numele autorului le mulţumesc.

OCTAV ONICESCU

Scanned with OKEN Scanner

O lămurire

Cartea de față, deși are la bază cursul predat studenților Școlii de Statistică și se adresează, prin urmare, în special acestora, cuprinde un câmp de probleme mai vast decât cel încadrat în lecțiile de școală. Am căutat ca prin aceasta să pun la dispoziția celor pe care metodele moderne ale statisticii matematice i-ar putea interesa o lucrare românească utilă și pe cât posibil mai completă. Majoritatea acestor metode sunt expuse în diverse periodice, unde cititorul, din cauza caracterului abstract al expunerii, le poate urmări mai greu. Pe deasupra, periodicele de specialitate lipsesc din cele mai bune biblioteci ale noastre.

Planul după care a fost întocmită această lucrare a avut un dublu scop.

Să dea, în primul rând, celor cu o pregătire matematică elementară posibilitatea utilizării celor mai noi procedee de cercetare; în al doilea, să pună pe cercetătorul teoretic în contact intim cu aplicațiile practice ale statisticii matematice, îndemnându-l astfel să-și aleagă problemele dintre acelea care într-adevăr pot prezenta o utilitate practică. În acest din urmă scop, am adăugat la sfârșitul lucrării un apendice care nu se adresează decât celor cu o pregătire matematică înaintată. Astfel se și explică faptul că amănuntele demonstrațiilor au fost uneori lăsate de o parte. Lectura apendicelui nu este câtuși de puțin necesară pentru o înțelegere completă a primelor patru părți; este absolut necesar ca cei ce o vor întreprinde să fi citit textul acestora din urmă.

Lucruri noi în cartea de față? Prea puține, în afară de ceea ce este mai întotdeauna nou într-o expunere a unor lucruri în majoritate cunoscute.

Am căutat să expun, pe cât mi-a permis spațiul, cele mai recente procedee ale statisticii matematice. Cadrul aplicațiilor a fost, pe de altă parte, astfel ales, încât cititorul să nu rămână cu o impresie trunchiată asupra posibilităților de interpretare numerică a fenomenelor colective. În ceea ce privește înseși aceste posibilități, ele au fost mereu considerate în lumina celor două clasice puncte de vedere: cel al școlii engleze - îndrăzneț și motivat de fertilitatea rezultatelor - și cel al școlii franceze și germane (al școlii „continentale", după cum obișnuiesc să o denumească anglo-saxonii) - descriptiv în mare măsură, obsedat de critica valabilității ipotezelor introduse și, tocmai din această cauză, triumfător, în multe puncte îndelung dezbătute, asupra celui dintâi. Astfel privite lucrurile, cartea poate că nu este nouă numai pentru literatura noastră.

Realizarea intențiilor mele n-ar fi fost posibilă fără un sprijin efectiv în această direcție.

Scanned with OKEN Scanner

Dl prof. O. Onicescu a îmbrăţişat fără şovăire ideea publicării în româneşte a unui tratat de statistică matematică. Paginile ce urmează au fost, capitol cu capitol, supuse criticii sale. Părţi importante din acestea se resimt după urma sugestiilor ce mi le-a făcut. Dar acest lucru nu este decât una din verigile lanţului început în toamna anului 1927, când, având sub formă de proiect o organizare modernă a statisticii în ţara noastră, mă îndemna să-mi îndrept studiile făcute în această direcţie. Să-mi fie permis să exprim, şi pe acestă cale, sentimentele de recunoştinţă şi de admiraţie ce le am pentru d-sa.

Dl dr. I. Teodorescu, directorul Institutului de Statistică Generală a Statului, a primit bucuros ca această lucrare să facă parte din „Biblioteca Institutului", iar dnii dr. L. Colescu, dr. S. Manuilă, Şerban Gheorghiu au subscris propunerii d-sale. Mulţumindu-le tuturor, îmi îndeplinesc mai mult decât o simţită datorie.

Trei dintre elevii Şcolii de Statistică: dra Amelia Suroiu, licenţiată în fizico-chimice, dnii I. Popa, licenţiat al Academiei de Comerţ, M. Rădulescu, licenţiat în matematici, au sacrificat spontan din timpul lor pentru a mă ajuta în munca aşa de migăloasă a controlului calculelor din numeroasele tabele, a revederii textului şi a corecturilor tipografice. Îi rog să primească viile mele mulţumiri camaradereşti.

Condiţiile de tipar excelente şi forma estetică sub care se prezintă lucrarea de faţă sunt datorate grijii deosebite pe care Imprimeria Naţională a depus-o în acest scop.

Autorul

Scanned with OKEN Scanner

PARTEA I

Scanned with OKEN Scanner

Capitolul I

Introducere

1. *Evoluția cuprinsului noțiunii de statistică.* Înțelesul ce s-a atribuit cuvântului *statistică* a evoluat de la o epocă la alta; azi chiar, tocmai din cauza acestei evoluții și a diferitelor curente din trecut, semnificația lui este interpretată în diferite chipuri.

Termenul *statistik* , derivat de la italianul *statista*, apare pentru prima oară în lucrarea lui GOTTFRIED ACHENWALL, intitulată *Abriss der Neuesten Staatswissenschaft der heutigen vornehmsten europäischen Reiche und Republiken* și publicată în 1749.

Însă la acea epocă, Școala lui Achenwall nu înțelegea prin acest termen decât un ansamblu de informații descriptive despre *lucrurile remarcabile* relative la colectivitatea ce se cheamă stat, expuse într-o anumită ordine. Stabilirea naturii informațiilor și a cadrului din care acestea făceau parte constituia pe atunci problema de căpetenie a statisticii.

Școlii engleze, ale cărei începuturi sunt reprezentate de lucrările lui PETTY, GRAUNT și HALLEY, îi revine meritul de a fi văzut, cea dintâi, încă din a doua jumătate a secolului al XVII-lea, avantajul informațiilor exprimate sub formă numerică. Ilustrările acestei idei nu lipsesc. GRAUNT pune în evidență faptul că nașterile masculine sunt mai numeroase decât cele feminine și stabilește raportul $\frac{14}{13}$ dintre numărul lor. Astronomul HALLEY calculează prima tablă de mortalitate.

Influența școlii engleze asupra celei germane n-a întârziat de a se produce, nu însă fără oarecare rezistență din partea ultimei și nu fără a da loc la polemici pe tema rigidității informațiilor exprimate în cifre, polemici care au dăinuit până târziu de tot, școala germană rămânând refractară metodelor întrebuințate de *aritmeticienii politici englezi.*

Între timp, scopurile finale urmărite prin statistică de către Școala lui Achenwall și-au găsit cristalizarea în apariția unor noi discipline științifice, ca: economia politică, sociologia, demografia, fapt care permite lui G.RUMELIN să afirme la 1874 că statistica nu-i decât o metodă. Totuși, divergențele părerilor asupra definiției ce trebuia adoptată pentru statistică au persistat. Ele au condus la plasarea de definiții telegrafice, ca: știința numărătorii, ba chiar a cifrelor, știința marilor numere, știința valorilor medii etc. Găsirea unei formule salvatoare fiind o chestiune multă vreme dezbătută, a contribuit în parte la răspândirea unei

Scanned with OKEN Scanner

opinii nu tocmai excelente despre statistică. Acest fapt împreună cu un număr de rezultate absurde obținute printr-o aplicare greșită, conștientă uneori, a principiilor metodei statistice au îngăduit lansarea cunoscutei glume: „Statistica este știința prin care se poate dovedi orice".

Azi există o tendință generală ca prin statistică să se înțeleagă metodele întrebuințate pentru culegerea, prezentarea și analizarea datelor numerice relative la fenomenele colective. Dezvoltarea pe care a luat-o această disciplină, astfel definită, a făcut necesară, mai ales în ultima decadă, crearea de noi termeni, ca statistica economică, financiară, politică, morală etc.

Se pare, în orice caz, că pentru a indica ansamblul metodelor de analizare a datelor, termenul de statistică matematică a fost general adoptat[1].

După YULE[2], statistica matematică este o serie de procedee «especially adapted to the elucidation of quantitative data affected by a multiplicity of causes»[3].

2. *Caracteristicile fenomenelor colective*. Fenomenele naturale, pe care diferitele discipline științifice caută să le studieze izolat sau în raporturile dintre ele, se împart, din punct de vedere al metodei întrebuințate pentru studiul lor, în două grupe bine distincte: fenomene tipice și fenomene colective. Fenomenele tipice sunt fenomenele care în condiții identice sau asemănătoare se prezintă sub aceeași formă unică. Experiențele făcute în domeniul acestor fenomene conduc, prin repetare, la același rezultat. Lumea neorganică este lumea fenomenelor tipice.

Cealaltă grupă de fenomene, nu mai puțin importante, sunt fenomenele care nu se pot reproduce în mod identic aproape niciodată. Astfel de fenomene se numesc fenomene colective. Din această categorie fac parte fenomenele sociale, biologice și, prin excepție la afirmația de mai sus, unele fenomene neorganice, cum ar fi cele meteorologice.

---

[1] Statistique mathématique *este termenul curent în literatura franceză. Literatura engleză are încă preferințe pentru cuvântul* statistics, *consecință a concepției matematice a școlii engleze de statistică. De curând însă, termenul* mathematical statistics *a început să apară din ce în ce mai des, tocmai pentru a evita ambiguitatea, deși este considerat de unii ca un pleonasm.*

[2] *U.G. Yule,* ***An Introduction to the Theory of Statistics****, London, Griffin, 1929, p.5.*

[3] *Această definiție lasă în afară partea referitoare la metodele de culegere și prezentare a datelor. Despre aceasta nici nu vom vorbi în rândurile acestei cărți. O expunere amănunțită a acestora ar rămâne fără folos pentru cititor, tehnica culegerii și prezentării datelor statistice nu se poate însuși decât în urma unei îndelungate experiențe pe tărâmul practic, iar o privire generală asupra chestiunii nu este necesară, metodele constând din principii destul de simple, foarte adesea indicate direct de simplul bun simț. De altfel, o înțelegere a diferitelor aspecte ale acestei tehnici nu poate fi deplină decât după o cunoaștere completă a procesului de prelucrare, la care, prin instrumentul statisticii matematice, vor fi supuse mai târziu datele.*

Scanned with OKEN Scanner

Diferența dintre fenomenele tipice și cele colective constă în numărul restrâns și precis de cauze care determină manifestările primelor și complexitatea celor de care depind fenomenele atipice. Cauzele care influențează ultima categorie de fenomene, în afară de faptul că sunt foarte variate și foarte numeroase, mai prezintă și caracterul esențial constând din faptul că aceeași cauză, adesea de minimă importanță, poate avea când un efect neglijabil, când un efect de preponderență capitală. Acest fapt dă fenomenelor colective înfățișarea întâmplării, a hazardului. Punând problema în mod invers, aceasta este definiția pe care H.POINCARÉ o dă hazardului[1].

Distincția aceasta, în raport cu modul cum fenomenele se prezintă studiului, nu este de altfel decât formală, complexitatea manifestărilor fenomenelor colective neținând decât de construcția spiritului nostru care nu poate urmări decât ceea ce este tipic într-o serie întreagă de experiențe. Totuși, metodele de cercetare și de descoperire a legilor care le guvernează trebuie să difere, tocmai din această pricină, de la o categorie de fenomene la alta.

Pentru studiul fenomenelor tipice, experiența de laborator unită cu metoda matematică sub forma clasică a acesteia este calea cea mai indicată de urmat. Însăși dezvoltarea metodei matematice, până la gradul pe care aceasta îl are azi, nu se datorește decât utilității pe care acest instrument îl prezintă pentru studiul fenomenelor tipice.

Caracterul special al fenomenelor colective nu permite, pentru cercetarea manifestărilor lor, experiența de laborator, care este bazată în special pe variația condițiilor cauzelor ce determină fenomenele. De multe ori chiar, experimentarea în domeniul fenomenelor colective este imposibilă. De aceea suntem nevoiți să ne folosim de manifestarea fenomenului așa cum ni-l oferă natura, fără ca noi să putem interveni.

În afară de aceasta, studiul manifestărilor fenomenelor colective nu se poate face decât în urma observării unui număr relativ mare de unități sau cazuri, din care să se desprindă «*l'absolu parmi le relatif, le typique parmi l'accidentel, le permanent parmi le passager*»[2]. Aceste aspecte ale lumii fenomenelor colective au făcut necesară crearea unei noi metode de cercetare, căreia să nu-i lipsească avantajul eficacității metodei matematice, dar care să fie adaptată în același timp formei aleatorii a manifestării acestor fenomene. Acestei cerințe îi corespunde statistica matematică, creată prin completarea noțiunilor matematicii clasice cu altele noi, specifice scopului propus, dintre care cele mai importante sunt noțiunea de *variabilă întâmplătoare*[3] și cea de *lege stocastică*.

3. *Legea stocastică*. După terminologia propusă de statisticianul suedez C.CHARLIER, legăturile dintre fenomene se exprimă prin *legi rigide* în cazul

---

[1] *H. Poincaré*, ***Science et méthode***, *Paris, Flamarion, articolul «Le hasard».*

[2] *A.Julin*, ***Principes de statistique théorique et appliquée***, *Paris, Marcel Rivière, 1921, vol.I, p.6.*

[3] Variable aléatoire ou fortuite, *termen introdus de școala matematică franceză. Termenul românesc este cel propus de dl prof.O. Onicescu.*

Scanned with OKEN Scanner

fenomenelor tipice, prin *legi stocastice* în cazul celor colective. Legile stocastice trebuie înţelese şi privite sub o lumină specială. Astfel, dacă ne referim la legea stocastică «Mens sana in corpore sano» nu trebuie să considerăm geniu orice atlet, după cum nu trebuie să vedem un idiot în orice pipernicit. Ar fi legitim să procedăm aşa dacă dezvoltarea trupească ar fi singura cauză determinantă a dezvoltării intelectului. Or, aceste două fenomene au un număr foarte mare de cauze determinante, din care numai o parte le este comună. Cauzele necomune, modifică parţial acţiunea celor dintâi. Această modificare variază de la un caz la altul şi astfel se explică de ce legea menţionată nu poate fi aplicată în acelaşi mod, ca, de pildă, legea rigidă «Electricităţile de acelaşi fel se resping, cele de fel contrar se atrag».

Cu toate acestea, modificarea acţiunii cauzelor comune de către celelalte cauze făcându-se când într-un sens, când într-altul, spiritul nostru, prin facultatea lui de abstractizare, desprinde din ansamblul de cazuri ceea ce este tipic numai şi astfel ajunge la formularea legii aşa cum am enunţat-o mai sus[1].

4. *Stabilitatea frecvenţelor*. Statistica matematică îşi propune, aşadar, să stabilească procedeele prin care aceste legi stocastice pot fi mai lesne descoperite, mai precis exprimate. În acest scop ea stabileşte legături între entităţile care-i sunt proprii, fără ca să le individualizeze. Aceste entităţi sunt considerate în cadrul metodologiei ca nişte noţiuni abstracte, la fel cum sunt noţiunile de punct, linie, număr, întrebuinţate în matematică. Rămâne ca adevărurile stabilite astfel *in abstracto* să fie apoi utilizate în domeniul practic ori de câte ori se poate stabili analogia între cazul teoretic şi cel oferit de realitate. Această legătură nu s-ar putea face - şi atunci eficacitatea metodei statistice ar deveni un cuvânt gol - dacă între teorie şi realitate n-ar exista dintr-un început un punct comun care să servească primei drept postulat sau ipoteză, piatra fundamentală a oricărei discipline ştiinţifice. Postulatul statisticii matematice este cunoscut sub numele de stabilitatea frecvenţelor. Forma sub care se enunţă de obicei acest postulat este următoarea:

«Dacă într-o serie de observaţii cuprinzând prima N_1 cazuri, a doua N_2, ultima N_p, *obţinute sub influenţa aceluiaşi complex de cauze*, f_1, f_2,..., f_p cazuri prezintă respectiv calitatea A, atunci raporturile:

$$\frac{f_1}{N_1}, \frac{f_2}{N_2}, \ldots, \frac{f_p}{N_p},$$

numite şi frecvenţe relative ale evenimentului A, nu diferă prea mult între ele dacă numărul cazurilor observate este suficient de mare».

Nu este departe timpul când acest adevăr, întrebuinţat de fizicieni sub formă mai completă, cunoscută sub numele de legea erorilor, era considerat de

[1] *Un exemplu simplu de joc de zaruri, pe care îl vom examina mai târziu, va scoate şi mai bine în relief acest caracter special al legilor stocastice (Cap.XI).*

Scanned with OKEN Scanner

aceştia drept adevăr matematic, în timp ce matematicienii credeau că ceilalţi deţin privilegiul învederării lui. Azi lucrurile sunt clarificate: stabilitatea frecvenţelor este pentru statistica matematică un postulat, iar pentru ştiinţă, în general, o lege *fizică* ce se poate constata, nu demonstra, în mod experimental. Acest fapt nu împiedică însă ca demonstraţii matematice ale stabilităţii frecvenţelor să apară din când în când. Unele din ele sunt adevărate capodopere de gândire matematică. Toate aceste demonstraţii pornesc de la ipoteze care, chiar dacă nu sunt prea complexe, au defectul de a nu fi naturale. Şi atunci e preferabil să postulăm adevărul de demonstrat însuşi.

Primul care a enunţat legea stabilităţii frecvenţelor a fost JACQUES BERNOULLI, în opera sa *Ars Conjectandi*, apărută în 1713. Faptul că Bernoulli ajunge la enunţarea ei în urma unor demonstraţii matematice a contribuit la admiterea fără discuţii a acestei legi, deşi demonstraţia dată este bazată pe ipoteze tot aşa de discutabile ca şi însăşi legea. Mirajul demonstraţiei matematice a servit, din fericire, la răspândirea aplicaţiilor acestui adevăr în diferite domenii mult mai repede decât dacă justificarea lui s-ar fi făcut pe altă cale.

5. *Stabilitatea fenomenelor colective*. Pentru structura legii stabilităţii frecvenţelor se obişnuieşte a se utiliza experienţa constând din aruncarea unui ban sau a unui zar, din scoaterea unei bile dintr-o urnă etc.

Aceste exemple au avantajul de a scoate în relief chintesenţa stabilităţii frecvenţelor. Ele prezintă însă marele dezavantaj de a lăsa complet în umbră o altă latură a acestei legi: experienţele menţionate, având întrucâtva un caracter artificial, nu pun în evidenţă faptul că acest adevăr este o interpretare a unei legi naturale, a cărei verificare se poate face la tot pasul. Această lege este legea fundamentală a fenomenelor colective şi ea nu ar mai constitui o realitate dacă aceste fenomene ar suferi variaţii continue, dacă legea stocastică observată azi n-ar mai fi valabilă mâine. Există, aşadar, şi o stabilitate a legilor stocastice, fapt care nu trebuie neglijat; ignorarea cuprinderii lui în postulatul statisticii matematice poate duce la rezultate greşite, care de cele mai multe ori sunt imputate metodei înseşi.

Tablele de mortalitate ne arată că din 100 000 de bărbaţi în vârstă de 30 de ani, 698 vor muri înainte de a împlini 31 de ani. Stabilitatea acestei cifre este aşa de demnă de încredere încât s-au şi găsit oameni care, organizându-se în societăţi de asigurare, să câştige bani din a paria contrariul. Nu constituie angajarea de capitaluri enorme în această direcţie şi mai ales prosperitatea acestor societăţi cea mai bună probă a stabilităţii fenomenelor colective?

Frecvenţa teoretică 698 din 100 000 nu este obţinută în realitate decât destul de rar, însă în majoritatea cazurilor numărul deceselor întâmplate în condiţiile specificate diferă prea puţin de 698. Explicaţia acestui fapt este simplă. Pentru o aceeaşi categorie de indivizi, ansamblul de cauze care determină structura vitală a acesteia dintr-un moment oarecare al vieţii este comun unui şir nu prea lung de generaţii. Sunt însă şi cauze care influenţează în mod special asupra fiecăreia dintre ele şi care determină pentru fiecare caz în parte diferenţa numărului deceselor faţă de valoarea teoretică sau, mai bine zis, medie.

Scanned with OKEN Scanner

Pentru ca aplicarea unei legi stocastice, ca cea de mai sus, să fie îndreptățită, trebuie ca populația ce servește drept obiect al aplicațiilor să fie suficient de numeroasă, astfel încât structura ei să fie ferită de particularități pe care o populație restrânsă mai întotdeauna le prezintă. Aplicarea unei legi stocastice unei populații ce nu prezintă toate variațiile de cazuri în proporția lor normală este tot așa de absurdă ca și tentativa de a prevedea momentul morții unui individ cu ajutorul unei table de mortalitate. *Legile stocastice n-au valoare decât aplicate maselor cu structură completă, și nicidecum cazurilor particulare.* Orice neglijare a acestui principiu poate duce la o infirmare aparentă a existenței legilor stocastice.

În ceea ce privește structura completă a maselor, ea este asigurată de numărul mare al cazurilor ce le alcătuiesc. Iată de ce stabilitatea frecvențelor este cunoscută și sub legea marilor numere.

Se mai poate întâmpla ca valorile numerice ce intră în exprimarea unei legi stocastice să sufere cu timpul variații, și aceasta, mai ales în cazul populațiilor în proces de evoluare. Acest fapt nu constituie o atingere adusă stabilității legii stocastice, stabilitatea acestora neexcluzând posibilitatea unei evoluții a fenomenelor colective, în special când este vorba de raporturile numerice ce exprimă legăturile dintre ele. Dimpotrivă, recalcularea din timp în timp a acestor *constante statistice* este o necesitate; ea ne poate prilejui constatarea unei astfel de evoluții, oglindită în variația acestor constante.

6. *Dispersie.* Un alt fapt căruia statistica matematică îi acordă o importanță deosebită este o aprofundare a modului cum se manifestă stabilitatea frecvențelor. Spuneam mai sus că valoarea medie sau teoretică a unei frecvențe nu este obținută în practică decât destul de rar, dar că în majoritatea cazurilor frecvențele obținute nu se abat prea mult de la prima valoare. Modul cum se grupează aceste abateri în raport cu mărimea lor constituie o chestiune cunoscută în statistica matematică sub numele de problema *dispersiei abaterilor.*

Importanța acesteia este capitală. Cunoașterea dispersiei fenomenului de mortalitate poate să dea actuarului o idee de mărimea câștigului sau a pierderii pe care o poate aștepta în urma diferitelor contracte ale asiguraților.

În același timp, dispersia abaterilor poate să constituie un factor de diferențiere a două fenomene având aceeași valoare medie. Un exemplu minunat ni-l oferă climatologia. Două regiuni, care au aceeași temperatură medie anuală, pot să se deosebească complet din punct de vedere al climei, dacă una din regiuni are o climă mediteraneană - și atunci prezintă abateri de temperatură foarte mici față de media anuală, - iar cealaltă - prin abaterile relativ mari față de aceeași temperatură medie - se află sub o climă continentală.

7. *Cercetătorii moderni.* Am arătat în cele de mai sus care sunt problemele de ordin general în vederea rezolvării cărora s-a creat instrumentul de cercetare care este statistica matematică. Astfel privit scopul acestei discipline, abia se mai poate recunoaște în el ideea pe care primii statisticieni o aveau despre statistică.

Scanned with OKEN Scanner

Concepţia actuală despre statistica matematică este relativ nouă. Un prim început spre noua doctrină îl face astronomul belgian ADOLPHE QUÉTELET. El este cel care pune în evidenţă pentru prima oară existenţa unei regularităţi printre fenomenele colective, adaptând pentru descoperirea acestor regularităţi metodele matematice ale calculului probabilităţilor. Ideile sale fundamentale sunt expuse în două lucrări mai importante: «*Sur l'homme et le développement de ses facultés ou Essai de physique sociale*»(1835) şi «*Lettres sur la théorie des probabilités appliquées aux sciences morales et politiques*» (1837).

Totuşi, şcolii engleze, şi mai ales şefului ei, profesorul KARL PEARSON, îi revine meritul de a fi precizat şi completat ideile lui Quételet, precum şi de a fi creat procedeele eficace ale metodei statisticii matematice, sub forma lor de azi. Imensa operă a lui Karl Pearson ar necesita volume întregi pentru a fi expusă. Lucrările sale se găsesc răspândite în mai toate periodicele ştiinţifice engleze, dar mai ales în jurnalul *Biometrika*, editat la început de Sir FRANCIS GALTON şi continuat apoi de el însuşi. Colecţia acestui periodic este un coran al statisticianului.

Din şcoala engleză, numele lui F.Y. EDGEWORTH se impune să fie, de asemenea, menţionat.

Paralel cu cercetările britanicilor, scandinavii, a căror figură reprezentativă în acest domeniu este C.V.L. CHARLIER, au dat naştere la o pleiadă de statisticieni care şi-au îndreptat activitatea mai ales în domeniul actuarial.

Alţi savanţi[1] şi alte şcoli, cum e cea italiană a lui CORRADO GINI, se ataşează, prin felul cum înţeleg să atace problemele de statistică matematică, fie şcolii engleze, mai ales, fie celei franceze.

Şcoala franceză, prin aproape unica ei figură mai reprezentativă, profesorul LUCIEN MARCH, a constituit un fel de frână a extinderii procedeelor statistice matematice în domenii unde aplicarea lor ar fi fost numai formală şi uneori chiar dăunătoare prestigiului adevărului ştiinţific. Lucrările lui Lucien March excelează printr-un foarte fin şi just spirit critic. El a fost acela care a introdus şi analizat noţiunea de *covariaţie*.

Procedeele statisticii matematice, o dată puse la punct de către cercetătorii pe care tocmai i-am pomenit, au fost şi sunt utilizate în scopuri cât mai multe şi mai variate. Nu este încă momentul să discutăm legitimitatea acestor aplicaţii; aceasta o vom face atunci când materialul expus ne va permite o analiză completă[2].

---

[1] *Ne gândim mai ales la doi statisticieni ruşi: A.Markoff şi A.Tschuprow.*

[2] *Cititorul doritor de a aprofunda mecanismul guvernator al fenomenelor colective se poate referi cu mult folos la studiul dlui prof.O. Onicescu,* ***Statistica şi sociologia****, Arhiva, anul IX, p. 498.*

Scanned with OKEN Scanner

Capitolul II

Variabile, unități, serii statistice

8. *Variabile*. Pentru studiul fenomenelor colective a fost necesară crearea unei entități care să oglindească în ea calitățile esențiale ale manifestării acestora. Acestei necesități îi corespunde noțiunea de *variabilă întâmplătoare*. Definiția matematică a variabilei întâmplătoare, deși cea mai completă, nu-și are locul aici. Noțiunea de variabilă în sine este prea simplă ca să mai fie nevoie de o analiză a ei. Ceea ce deosebește fundamental noțiunea de variabilă întâmplătoare de cea de variabilă în sensul clasic este faptul că variațiile primei, fiind supuse unui ansamblu prea complex de cauze, nu pot fi determinate înainte de observarea efectivă a fenomenului.

Exemple de fenomene ale căror manifestări pot fi reprezentate prin variabile întâmplătoare ne stau ușor la dispoziție mai în toate domeniile. Astfel, statura unei persoane, salariul pe care-l primește săptămânal un lucrător, numărul falimentelor întâmplate în cursul unei anumite perioade de timp și în limitele unei unități administrative sunt ilustrații comune ale noțiunii de variabilă întâmplătoare.

Variabilele pot fi împărțite după modul cum se manifestă variația lor, în două clase:

1. Variabile discontinue;
2. Variabile continue.

Variația variabilelor discontinue se manifestă prin salturi; astfel de variabile *nu pot lua decât anumite valori* pe scara lor de variație firească. Acesta este cazul variabilei ce reprezintă numărul copiilor dintr-o familie și care poate fi 0,1,2,3,..., nici o familie însă nu poate avea fracțiuni de copii. Aceeași constatare se poate face și pentru numărul comunelor dintre granițele unei țări etc.

A doua categorie de variabile cuprinde pe acelea care pot lua *absolut orice valoare* din scara lor de variație. Astfel de variabile reprezintă, de pildă, greutatea unei persoane, tonajul unui vas, forța motrice a unei instalații etc. În practică însă, continuitatea acestor variabile este îngrădită de limita preciziei cu care noi putem efectua o măsurătoare: înălțimea diferiților indivizi alcătuind o populație ce se măsoară în centimetri sau milimetri, cel mult. Alteori, variabile

Scanned with OKEN Scanner

continue apar ca variabile discontinue, din cauza faptului că măsurătoarea nu se face decât cu o aproximaţie foarte grosolană, o precizie prea mare nefiind întotdeauna utilă pentru scopul ce se are în vedere. Tonajul unui vas se măsoară în tone, deşi precizia în măsurarea capacităţii de încărcare poate să treacă cu mult dincolo de această limită.

Multe din variabilele discontinue pot fi considerate, la rândul lor, ca nişte variabile continue. Populaţia unei ţări face parte din această categorie. În acest caz, variabila are indiscutabil caracterul unei variabile discontinue, rezultatul unui recensământ este întotdeauna exprimat printr-un număr întreg. Totuşi, salturile discontinuităţii sunt aşa de mici în raport cu ordinul de mărime al variabilei, încât aceasta poate fi considerată, fără eroare apreciabilă, ca o variabilă continuă. Acelaşi lucru se întâmplă cu variabilele ale căror valori sunt exprimate prin raportul a două numere întregi, cum este producţia medie lunară a unei fabrici, venitul pe cap de locuitor etc., deşi un astfel de raport nu poate lua absolut orice valoare[1].

9. *Unităţi statistice*. Operaţia de măsurare sau de numărare necesită fixarea în prealabil a unei unităţi. Unităţile întrebuinţate în statistică cer o atenţie specială din cauza varietăţii aspectelor sub care se poate prezenta acelaşi atribut în fapt. Definiţia unei unităţi, când aceasta se prezintă în mod natural, nu comportă aproape nici o dificultate. Unităţile construite de statistică sau cele ce sunt în legătură cu instituţii de ordin social, economic, moral nu se conturează tot aşa de uşor ca în cazul precedent. Deşi noţiunea de emigrant este destul de clară în mintea oricui, definiţia ei, adoptată în statistică, poate să ridice chestiuni din cele mai complicate. Această definiţie trebuie să permită în primul rând recunoaşterea efectivă a emigranţilor, fapt fără de care o culegere de date în această privinţă nu este posibilă. Definiţia emigrantului, aşa după cum a fost adoptată de către birourile de emigraţie din diferitele ţări, este următoarea: se numeşte emigrant orice cetăţean ce părăseşte ţara sa printr-un port şi într-un grup condus de o călăuză. Se vede cât de diferită este această definiţie de noţiunea comună de emigrant. Aceasta are însă două calităţi esenţiale din punct de vedere al metodei statistice:

1. Evită aproape orice ambiguitate şi reduce, prin urmare, la minimum influenţa ecuaţiei personale a funcţionarului care face înregistrarea statistică;

2. Face posibilă această înregistrare. Într-adevăr, dacă definiţia ar fi cuprins şi pe emigrantul care părăseşte ţara singur şi prin alte mijloace de transport, înregistrarea multora din emigranţi ar fi devenit o imposibilitate de fapt.

---

[1] *Într-adevăr, nu se poate găsi un raport care să fie egal cu* $\sqrt{2}$, *de pildă. Se dovedeşte în matematici că aceste exemple sunt mult mai numeroase decât par. Ansamblul numerelor fracţionare este departe de a fi continuu. Este totuşi, cu un termen tehnic, dens peste tot; densitatea aceasta ne permite să-l considerăm continuu în statistică.*

Scanned with OKEN Scanner

Acestea sunt două condiţii esenţiale pe care orice definiţie de unitate statistică trebuie să le îndeplinească. Oricât de simple ar părea la prima vedere, aceste două condiţii nu sunt din cele mai uşor de satisfăcut. Un exemplu în această direcţie îl constituie definiţia statistică a «omului sărac», definiţie care n-a putut să corespundă cerinţelor menţionate, deşi necesitatea ei este evidentă pentru un studiu al pauperismului dintr-o ţară. De aceea, pauperismul este studiat azi prin indici indirecţi, ca numărul ajutoarelor distribuite gratuit de către societăţile de binefacere, numărul de contravenienţi arestaţi pentru cerşit etc.

Un alt exemplu tipic în această direcţie îl constituie definiţia copilului născut mort. Dacă în exemplele precedente s-ar putea ca însuşi individul considerat să nu poată răspunde dacă este sau nu un emigrant sau dacă este sau nu un sărac, în cazul unui născut se poate preciza printr-o autopsie a cadavrului dacă avem într-adevăr de a face cu un copil născut mort sau dacă acesta a murit în timpul naşterii ori cu câteva momente în urmă. Trebuie să recunoaştem că nu putem obliga oficiile demografice să facă această autopsie ori de câte ori este cazul. Şi atunci este nevoie să adoptăm un criteriu în baza căruia să facem posibilă înregistrarea statistică a născuţilor morţi, fără nici o ambiguitate. Se poate adopta, de pildă, următoarea definiţie: Este considerat născut mort orice copil născut după o perioadă de uterogestaţie mai mare de 180 de zile şi care este mort în momentul declaraţiei naşterii la oficiul de stare civilă. Aceasta este definiţia care s-a adoptat de către statistica Franţei. Cum declaraţia la oficiul stării civile trebuie să fie făcută, după lege, în termen de cel mult trei zile de la naştere, definiţia de mai sus va conduce de fapt la o sporire a numărului de născuţi morţi biologiceşte cu acel al născuţilor vii, morţi în primele trei zile de viaţă. Nu trebuie însă să privim acest lucru ca un prea mare inconvenient, căci ceea ce ne interesează, în special, în legătură cu numărul de născuţi morţi este evoluţia acestuia în timp. Or, numărul de născuţi vii, morţi în primele trei zile de la naştere, pare a varia proporţional cu cel al născuţilor morţi biologic şi atunci variaţia acestuia din urmă va fi destul de exact redată de variaţia numărului de născuţi morţi după definiţia statistică.

Statisticianul care-şi propune analizarea manifestărilor unui fenomen colectiv, oglindite într-un ansamblu de date statistice, are datoria ca înainte de toate să cunoască precis definiţia unităţii statistice ce a servit la culegerea acestor date, în primul rând, pentru ca să evite orice apreciere - în plus sau în minus - asupra intensităţii fenomenului considerat, în al doilea rând, ca să nu ajungă la concluzii greşite prin compararea datelor statistice culese cu ajutorul a două unităţi diferite. Pentru ca scopul publicaţiilor statistice să fie efectiv atins, trebuie, aşadar, ca aceste publicaţii să fie însoţite de explicaţii complementare, atât asupra unităţii statistice respective, cât şi asupra mijloacelor întrebuinţate pentru culegerea datelor. Din nefericire, acest deziderat nu este întotdeauna respectat, uneori dintr-o neînţelegere regretabilă a lucrurilor, alteori pentru a se ascunde anumite defecte de tehnică. Aici stă, poate, unicul neajuns al statisticii, dar care, atunci când există, anulează toate celelalte avantaje ale metodei.

Scanned with OKEN Scanner

10. *Serii statistice*. Rezultatele unei înregistrări statistice se prezintă sub forma unei serii. Prezentarea ordonată a datelor privitoare la manifestările unui fenomen colectiv constituie seria statistică. Aceste manifestări pot fi raportate la un anumit atribut al lor sau la diferitele aspecte sub care se pot înfăţişa.

Diferenţierea manifestărilor se poate face, în primul rând, după locul unde s-a produs fenomenul, în al doilea rând, în raport cu timpul când s-a făcut observarea lui, în fine, în raport cu un atribut oarecare, altul decât timpul sau spaţiul. De aici decurg trei tipuri de serii statistice, care, bineînţeles, pot da naştere, prin combinări, şi la alte forme.

Seria statistică, aşa cum o cunoaştem cu toţii din orice publicaţie de date statistice, se prezintă ca o corespondenţă, o legătură între două şiruri de valori. Primul şir constituie scara atributului în raport cu care s-a făcut înregistrarea statistică, al doilea depinde de felul cum este alcătuit cel dintâi şi este determinat prin observarea fenomenului corespunzător. O serie statistică poate fi deci privită ca exprimarea unei variabile în raport cu variaţia alteia. Astfel, în tabelul 1, avem exprimată temperatura medie anuală a câtorva localităţi din România în raport cu variaţia altitudinii.

TABELUL 1

Altitudinea şi temperatura medie anuală a câtorva localităţi din România

Localitatea	Altitudinea în metri	Temperatura medie anuală
Sinaia	860	5,1
Braşov	570	7,8
Cluj	365	7,7
Cernăuţi	243	6,8
Bucureşti	82	11,1
Constanţa	4	10,3

Date extrase din ***Anuarul statistic al României****, 1930, p. 12-15*[1].

Cu un termen matematic, prima variabilă (altitudinea) se numeşte *variabilă independentă*, cea de-a doua (temperatura medie anuală), funcţie. Valorile variabilei independente constituie elementul de diferenţiere a diferitelor cazuri observate. Trebuie să avem în vedere că uneori variabila independentă poate să nu fie exprimată numeric, dar acest fapt nu micşorează întru nimic

[1] *Pentru ilustrarea procedeelor de analiză expuse în aceste pagini, am căutat, pe cât lucrul a fost cu putinţă, să mă servesc de serii statistice referitoare la ţara noastră. Este înţeles de la sine că numai datele neprelucrate au fost împrumutate de la sursele pe care întotdeauna le-am citat.*

Scanned with OKEN Scanner

constatările de mai sus. În tabelul 2 variabila independentă este sexul, care nu poate fi exprimat printr-un număr decât în cazul când se stabilește o convenție în acest sens; acest atribut constituie totuși elementul de diferențiere a numărului de elevi.

TABELUL 2

Numărul elevilor înscriși în școlile rurale din Vechiul Regat, în anul 1929/30

Sexul	Numărul elevilor
Băieți	459 759
Fete	390 878
Total	**850 637**

Extras din ***Anuarul statistic al României****, 1930, p. 339.*

11. *Distribuții în spațiu.* Seriile statistice în care elementul de diferențiere a manifestărilor fenomenului corespunzător este locul în care acestea s-au produs poartă numele de serii sau distribuții în spațiu. Seriile statistice privind presiunea atmosferică din diferitele stații de observație, nivelul mediu al costului vieții în diferite orașe, activitatea diferitelor porturi ale unei țări, densitatea populației din diferite împărțiri administrative etc. sunt exemple tipice de distribuții în spațiu.

De obicei, aceste serii se prezintă sub forma unui șir de date; observațiile care au condus la aceste date trebuie să fie făcute în toate punctele de observație în același interval de timp. Epoca la care s-au făcut observațiile constituie un atribut al seriei în spațiu și, ca atare, trebuie menționată o dată cu celelalte amănunte privitoare la culegerea datelor.

În ceea ce privește chestiunea prezentării datelor alcătuind o serie în spațiu, orice ordine logică poate fi adoptată. Astfel, în tabelul 3 datele au fost înșirate în ordinea descrescătoare a mărimii lor.

Perioada de observație în cazul de față a fost 1 ianuarie - 31 decembrie 1930; variabila independentă este poziția geografică a localității, iar funcția, nivelul costului vieții. Unitatea statistică corespunzătoare acestui tabel este o unitate abstractă: nivelul mediu al prețurilor din 1913[1].

Un alt mod destul de întrebuințat pentru prezentarea seriilor în spațiu este ordonarea datelor după poziția geografică a unității administrative corespunzătoare. În publicațiile noastre statistice, mai ales când este vorba de unități administrative mai mari, se obișnuiește să se facă clasificarea datelor după provincii istorice. Diferitele provincii care formează România de azi, găsindu-se până la 1918 sub regimuri diferite, au o structură deosebită; acest mod de

[1] *Vezi capitolul* Numere indice.

Scanned with OKEN Scanner

prezentare a datelor pe provincii istorice are avantajul de a pune în evidență această diferență și de a permite eventual constatarea atenuării ei cu timpul.

TABELUL 3

Indicele de cost al vieții în principalele orașe din România, în cursul anului 1930

(Baza 1913 = 100)

Orașe	Indice
București	4 724
Cluj	4 484
Cernăuți	4 413
Constanța	4 381
Ploiești	4 380
Iași	4 370
Chișinău	4 359
Brașov	4 340
Galați	4 300
Arad	4 282
Sibiu	4 278
Brăila	4 270
Oradea	4 264
Timișoara	4 238
Tg.-Mureș	4 160
Craiova	4 096

*Extras din **Statistica prețurilor pe anul 1930**, p.6.*

12. *Serii în timp.* O a doua formă sub care se pot prezenta manifestările unui fenomen este un șir de date corespunzătoare acestora, obținute la diferite intervale de timp. O astfel de serie statistică poartă numele de serie în timp.

Unitatea cea mai des adoptată pentru seriile în timp este anul civil. Adoptarea unei unități de timp mai scurtă decât anul este îndreptățită mai ales în cazurile când fenomenul prezintă variații determinate de anotimp, de poziția zilei în săptămână, a orei în intervalul unei zile. De foarte multe ori însă, unitatea statistică ne este indicată din afară. Producția agricolă nu poate fi urmărită în timp decât anual, deoarece recoltarea ei se face, cu mici variații, la o anumită epocă a anului, pentru o aceeași regiune agricolă.

Tabelul 4 ne arată forma sub care se prezintă de obicei o serie în timp. Variabila independentă este aici timpul, iar unitatea pentru măsura lui, anul. Producția de țiței, în cantitate sau valoare, are rol de funcție.

Scanned with OKEN Scanner

TABELUL 4

Producția de țiței brut din România pentru anii 1919-1930

Anii	Producția în tone	Valoarea în lei
1919	855 542	143 589 961
1920	1 108 924	739 645 795
1921	1 168 414	1 021 296 259
1922	1 372 905	1 875 861 426
1923	1 512 302	3 756 254 774
1924	1 860 471	4 584 263 454
1925	2 316 979	5 750 678 220
1926	3 244 415	7 392 769 540
1927	3 669 354	6 106 705 692
1928	4 282 377	7 345 923 066
1929	4 836 974	7 441 204 190
1930	5 792 311	6 552 118 582

*Extras din **Statistica minieră pe anul 1930**, p. V.*

Sunt cazuri însă în care anul luat ca unitate de măsură pentru timp ascunde anumite caractere ale fenomenelor, caractere ce nu sunt câtuși de puțin neglijabile pentru o cunoaștere completă a fenomenului studiat.

Tabelul 5 ne arată că există o diferență indiscutabilă între numărul căsătoriilor contractate în diferitele luni ale anului, diferență ce-și găsește explicația în existența diferitelor obiceiuri și reguli religioase în legătură cu celebrarea căsătoriei.

TABELUL 5

Numărul căsătoriilor contractate în cursul anilor 1904-1905, în România, repartizate pe luni

Lunile	1904		1905	
	Căsătorii	%	Căsătorii	%
Ianuarie	15 923	30,7	6 600	12,9
Februarie	5 786	11,1	11 314	22,1
Martie	434	0,8	438	0,9
Aprilie	1 721	3,4	741	1,4
Mai	2 176	4,2	1 221	2,4
Iunie	837	1,6	1 233	2,4
Iulie	1 381	2,6	1 225	2,4

Scanned with OKEN Scanner

(continuare tab. 5)

Lunile	1904		1905	
	Căsătorii	%	Căsătorii	%
August	1 482	2,8	1 330	2,6
Septembrie	3 300	6,4	2 664	5,2
Octombrie	7 659	14,8	7 907	15,4
Noiembrie	7 786	15,0	10 551	20,6
Decembrie	3 429	6,6	6 021	11,7
Total	**51 914**	**100,0**	**51 245**	**100,0**

*Extras din **Mişcarea populaţiei României pe anii 1904-1905**, p.7 şi 87.*

Alteori, poate să fie util a urmări mersul zilnic al unui fenomen, fie că variaţiile zilnice ale manifestărilor acestuia prezintă o importanţă intrinsecă, fie că aceste manifestări sunt influenţate de activitatea hebdomadar periodică a omului. Temperatura unui bolnav e înregistrată zilnic, fiind în general unul din simptomele caracteristice ale mersului unei maladii. Numărul de accidente de muncă sau de circulaţie, de indivizi găsiţi în stare de ebrietate diferă după poziţia zilei în săptămână (tabelul 6).

TABELUL 6

Distribuţia accidentelor de circulaţie în raport cu zilele săptămânii

Ziua	Numărul accidentelor	Procente
Duminică	1 563	10,03
Luni	2 386	15,31
Marţi	2 305	15,05
Miercuri	2 346	14,79
Joi	2 214	14,68
Vineri	2 288	14,20
Sâmbătă	2 486	15,94
Total	**15 588**	**100,00**

*Extras din **Statistique des accidents de roulage**, Belgia, 1929, p.18.*

Societăţile de transport, de electricitate, de telefoane au tot interesul să meargă mult mai departe cu analiza mersului în timp al fenomenelor care le privesc, pentru ca astfel să-şi aranjeze numărul funcţionarilor, vehiculelor de transport etc. în acord cu cererea publicului. Tabelul 7 pune în evidenţă faptul că în cursul unei zile există un ritm destul de accentuat al numărului convorbirilor telefonice.

Scanned with OKEN Scanner

TABELUL 7

Distribuția apelurilor telefonice în raport cu ora din zi

Ora	Apeluri telefonice la 10 000
5- 6	44
6- 7	72
7- 8	281
8- 9	758
9-10	1 081
10-11	1 102
11-12	1 047
12-13	838
13-14	399
14-15	335
15-16	595
16-17	819
17-18	860
18-19	754
19-20	475
20-21	296
21-22	167
22-23	56
23-24	21
	10 000

Calculat după datele obținute prin bunăvoința Serviciului statistic al Societății Române de Telefoane.

13. *Serii în timp cumulate.* Seriile în timp se mai pot prezenta și sub forma numită *cumulativă.* Termenii unei serii cumulative se formează din termenii unei serii simple, prin cumulare (adunare) succesivă în ordinea lor cronologică, astfel încât un termen din seria cumulativă reprezintă totalul termenilor din seria simplă ce se găsesc înaintea acestuia sau la același nivel cu el. Bineînțeles că, după însăși definiția aceasta a seriei cumulative, seria simplă se poate deduce prin scăderea termenilor seriei cumulative în timp - fiecare din cel următor - cu excepția primului termen care este întotdeauna comun ambelor serii. Tabelul 8 oferă un exemplu de serie în timp sub forma simplă, alături de seria cumulativă corespunzătoare; pe acest exemplu se pot verifica lesne toate proprietățile menționate mai sus.

Seria în timp sub forma cumulativă își găsește explicații mai ales în urmărirea realizării unui plan de activitate dinainte stabilit, cum ar fi încasările bugetare, subscrierile pentru emisiunea unor rente etc. Dar asupra acestei chestiuni vom mai avea ocazia să revenim.

Scanned with OKEN Scanner

TABELUL 8

Exportul de petrol brut al României în cursul anului 1930

(Cantităţi în tone, 1 tonă = 1000 kg)

Luna	Producţia	
	lunară	cumulată de la 1 ianuarie
Ianuarie	4 894	4 894
Februarie	1 377	6 271
Martie	1 779	8 050
Aprilie	1 002	9 052
Mai	3 196	12 248
Iunie	6 569	18 817
Iulie	1 906	20 723
August	1 035	21 758
Septembrie	2 352	24 110
Octombrie	4 767	28 877
Noiembrie	35 416	64 293
Decembrie	5 068	69 361

*Extras din **Statistica minieră a României pe anul 1930**, p. 104-105.*

14. *Distribuţii de frecvenţe*. O ultimă categorie de serii este constituită de distribuţiile de frecvenţă. Acestea sunt considerate de majoritatea statisticienilor ca fiind cele mai importante dintre seriile statistice.

Distribuţiile de frecvenţă sunt serii statistice în care variabila independentă este reprezentată de un atribut oarecare - altul decât spaţiul sau timpul - cum ar fi, de exemplu, înălţimea unui individ, numărul de petale ale unei flori, venitul recunoscut al unei persoane, culoarea ochilor etc. Variabila dependentă, funcţia, este în acest caz frecvenţa, adică numărul de cazuri observate ce se prezintă sub forma corespunzătoare a atributului determinat.

Atributul în raport cu care se stabileşte distribuţia de frecvenţă poate fi *calitativ*, şi atunci distribuţia poartă numele de *homogradă*, sau *cantitativ*, când distribuţia se numeşte *heterogradă* (terminologie introdusă de C.CHARLIER).

O distribuţie a unei populaţii în raport cu vârsta diferiţilor indivizi ce o alcătuiesc este o distribuţie heterogradă, vârsta fiind un atribut cantitativ. Distribuţia aceleiaşi populaţii în raport cu starea civilă este, din contră, o distribuţie homogradă, căci starea civilă este o calitate, ea nu poate fi exprimată numeric.

Tabelul 9 constituie un exemplu de distribuţie homogradă - distribuţia numărului crimelor în raport cu felul cum acestea sunt încadrate în Codul penal.

Scanned with OKEN Scanner

TABELUL 9

Numărul și felul crimelor comise în Vechiul Regat în cursul anului 1928

Felul crimei	Numărul crimelor
Tâlhării	255
Tentative de omor, omoruri	151
Lovituri cauzatoare de moarte	146
Pruncucideri	37
Incendieri	13
Provocări de avorturi	6
	608

*Extras din **Statistica judiciară a României pe anii 1925-1928**, p. 21.*

Aici este locul să observăm că distribuția în spațiu se apropie mult, prin forma sub care se prezintă, de distribuția homogradă; în definitiv, dacă nu ținem seama de determinarea unei localități prin coordonatele ei geografice, ci de distingerea acesteia prin numele pe care-l poartă, suntem evident în fața unei distribuții în raport cu un atribut ce nu este cantitativ.

Tot o distribuție homogradă este aceea înfățișată de tabelul 10.

TABELUL 10

Producția de cereale a României în anul 1930

Cereale	Cantități în chintale
Porumb	45 199 201
Grâu	35 590 379
Orz	23 713 061
Ovăz	11 565 361
Secară	4 645 292
Mei	788 412
Mături	124 040
Hrișcă	17 653

*Extras din **Anuarul statistic al României pe anul 1930**, p.50-51.*

Deși această serie nu se prezintă, în formă, ca o distribuție de frecvențe, trebuie totuși să o privim ca atare. Chintalele nu sunt altceva decât unități colective, formate din mai multe boabe sau molecule, dacă vrem; or, ar fi să împingem lucrurile prea departe dacă am căuta să numărăm frecvențele boabelor sau ale moleculelor corespunzătoare fiecărei clase din distribuția considerată. Iată de ce seriile de felul acesteia trebuie încadrate în categoria distribuțiilor homograde.

Scanned with OKEN Scanner

De obicei, o distribuție homogradă se prezintă în ordinea descrescătoare a frecvențelor. Se poate însă întâmpla ca între diferitele clase să se poată stabili o gradație ori o ordine naturală sau conceptuală, cum ar fi cazul unei distribuții în raport cu culoarea ochilor, a capacității intelectuale etc. În acest din urmă caz, e preferabil să prezentăm distribuția în ordinea indicată de însăși această scară graduală.

TABELUL 11

Distribuția cetățenilor din New York, având o profesie liberă, în raport cu venitul lor anual

Venitul în dolari		Frecvența veniturilor
mai mare sau egal cu	mai mic decât	
-	1 000	14
1 000	2 000	675
2 000	3 000	1 691
3 000	4 000	1 304
4 000	5 000	1 081
5 000	6 000	600
6 000	7 000	482
7 000	8 000	389
8 000	9 000	302
9 000	10 000	211
10 000	11 000	194
11 000	12 000	151
12 000	13 000	127
13 000	14 000	130
14 000	15 000	100
15 000	20 000	370
20 000	25 000	200
25 000	30 000	125
30 000	40 000	179
40 000	50 000	88
50 000	60 000	50
60 000	70 000	44
70 000	80 000	26
80 000	90 000	15
90 000	100 000	11
100 000	-	23
	-	**8 582**

Extras din ***Annual Report of the New-York State Tax Commission****, 1922, p. 508-509.*

Scanned with OKEN Scanner

Tabelul 11 ne înfățișează o distribuție heterogradă. Cazurile înregistrate se diferențiază după nivelul la care se ridică venitul cetățenilor din orașul New York, având o profesie liberă recunoscută. Atributul în raport cu care s-a făcut distribuția este cantitativ, deci măsurabil.

15. *Intervalele scării de distribuție.* O chestiune de primă importanță pentru studiul distribuțiilor heterograde este modul cum se fixează intervalele ce alcătuiesc scara clasificării elementelor observate. Acestă importanță nu poate fi îndeajuns de evidențiată; de felul cum este alcătuită scara unei distribuții heterograde depinde modul în care este pus în evidență caracterul esențial al acesteia. Uneori, o scară greșit aleasă poate schimba complet aspectul distribuției frecvențelor. Vom insista mai mult asupra acestui fapt în paragrafele următoare, când, cu ajutorul reprezentării grafice, ideile acestea vor putea fi mai complet expuse și deci mai lesne urmărite. Deocamdată, ne vom mărgini să atragem atenția asupra faptului care deosebește scările adoptate în tabelele 11 și 12. Scara ultimului tabel este formată din intervale egale, pe când în tabelul 11 intervalele variază ca mărime după poziția lor în scară.

TABELUL 12

Distribuția bărbaților căsătoriți în raport cu vârsta

Vârsta	Frecvențe relative
15-20	2
20-25	38
25-30	96
30-35	123
35-40	132
40-45	129
45-50	123
50-55	110
55-60	88
60-65	70
65-70	46
70-75	26
75-80	12
80-85	4
85-90	1
	1 000

Date calculate după cele din ***Récensement de la population dans la République Tchécoslovaque****, 1920, Tome I-er, p. 254.*

Este de recomandat ca orice distribuție de frecvență să fie făcută pe o scară cu intervale egale, deoarece numai în acest caz frecvențele diferitelor

Scanned with OKEN Scanner

clase sunt direct comparabile între ele; or, diferența între mărimea acestor frecvențe constituie elementul cel mai caracteristic al unei distribuții. În cazul tabelului 11, frecvența corespunzătoare intervalului 14 000-15 000 nu este direct comparabilă cu frecvența clasei imediat următoare, 15 000-20 000. Prima are valoarea 100, iar cea de-a doua, 370; saltul acesta este datorat faptului că ultimul interval este de cinci ori mai mare decât primul și nu oglindește un aspect real al distribuției.

Cu toate acestea, există statistici publicate sub forma de distribuții heterograde pe o scară cu intervale neegale. Să trecem cu vederea cazurile când acest lucru se întâmplă dintr-o neînțelegere regretabilă a scopului urmărit de statistică și să ne oprim asupra celor în care adoptarea unei astfel de scări nu poate fi evitată. Să considerăm, de pildă, distribuția comunelor urbane din România în raport cu numărul locuitorilor. Comuna urbană care are cea mai mică populație (871 locuitori) este Carmen-Sylva, cea care are populația cea mai numeroasă, municipiul București (631 288 locuitori). Diferența așa de mare între populația acestor două comune ne împiedică să alegem un interval prea mic pentru scara distribuției heterograde corespunzătoare. Chiar cifra de 25 000 este prea mică pentru mărimea intervalelor în cazul de față, căci distribuția ar cuprinde atunci 26 de intervale, dintre care numai primele 5 și ultimul ar prezenta frecvențe diferite de zero (această afirmație decurge imediat din simpla inspecție a tabelului 13 A). Nici adoptarea unei scări având intervalele egale cu 50 000 nu ar înlătura acest aspect «șchiop» al distribuției. Mergând mai departe și admițând pentru mărimea intervalelor cifra de 100 000 locuitori, obținem distribuția:

Între	0	și	99999 locuitori	167	comune
"	100 000	"	199 999 "	4	"
"	200 000	"	299 999 "	nici o comună	
"	300 000	"	399 999 "	"	"
"	400 000	"	499 999 "	"	"
"	500 000	"	599 999 "	"	"
"	600 000	"	699 999 "	1	comună

Or, această distribuție nu ne spune altceva decât că aproape totalitatea comunelor urbane din România au o populație mai mică decât 100 000 de locuitori. A aștepta numai atât după urma studiului distribuțiilor heterograde ar fi prea puțin; statistica matematică ar fi renunțat de mult la acest mijloc de cunoaștere a fenomenelor colective, dacă distribuțiile heterograde, la care acestea dau naștere, n-ar fi permis o pătrundere mai adâncă a lucrurilor. Este, așadar, necesar ca uneori să se renunțe la egalitatea intervalelor sau, dacă nu, să se modifice întrucâtva seria corespunzătoare.

Scanned with OKEN Scanner

TABELUL 13 A

Distribuția comunelor urbane din România în raport cu numărul locuitorilor[1]

(fără municipiul București)

Numărul locuitorilor	Frecvența comunelor
0 - 4 999	29
5 000 - 9 999	57
10 000 - 14 999	28
15 000 - 19 999	21
20 000 - 24 999	6
25 000 - 29 999	5
30 000 - 34 999	8
35 000 - 39 999	2
40 000 - 44 999	-
45 000 - 49 999	2
50 000 - 54 999	-
55 000 - 59 999	2
60 000 - 64 999	1
65 000 - 69 999	1
70 000 - 74 999	-
75 000 - 79 999	2
80 000 - 84 999	1
85 000 - 89 999	-
90 000 - 94 999	1
95 000 - 99 999	1
100 000 - 104 999	2
105 000 - 109 999	-
110 000 - 114 999	1
115 000 - 119 999	1
	171

Dacă admitem acest din urmă procedeu și lăsăm laoparte capitala țării care are o poziție singulară în ansamblul comunelor urbane, obținem pe o scară cu intervalele egale cu 5 000 de locuitori distribuția înfățișată în tabelul 13A. Dar și această formă de distribuție are dezavantajul de a prezenta clase pentru care distribuția corespunzătoare este nulă. De altfel, către sfârșitul ei, distribuția are aproape forma unei simple enumerări; și nici acesta nu poate fi scopul urmărit prin clasificarea în distribuție heterogradă. Se pot evita aceste inconveniente

[1] *Pentru formarea acestei distribuții, m-am servit de datele luate din* ***Indicatorul statistic al satelor și unităților administrative din România*** *(publicația Comisiei Recensământului General al Populației, 1930).*

numai dacă se recurge la o scară în care intervalele să fie mai mari către sfârşitul distribuţiei. Şi iată cum, în mod logic, ajungem la necesitatea de a utiliza uneori distribuţii pe intervale neegale.

O dată ce pentru distribuţia considerată este necesar să se renunţe la avantajul intervalelor egale, e bine să se profite de acest fapt pentru ca noua scară să fie astfel aleasă încât să permită o cunoaştere mai precisă a distribuţiei comunelor mici, ceea ce se poate face dacă adoptăm intervalul de 1 000 locuitori pentru primele clase ale distribuţiei. Astfel ajungem la distribuţia din tabelul 13B, care cuprinde şi municipiul Bucureşti.

TABELUL 13 B

Distribuţia comunelor urbane din România în raport cu numărul locuitorilor

Numărul locuitorilor	Frecvenţa comunelor
Sub 1 000	2
1 000 - 1 999	3
2 000 - 2 999	4
3 000 - 3 999	9
4 000 - 4 999	11
5 000 - 9 999	57
10 000 - 14 999	28
15 000 - 19 999	21
20 000 - 24 999	6
25 000 - 29 999	5
30 000 - 39 999	10
40 000 - 49 999	2
50 000 - 99 999	9
100 000 - 199 999	4
Peste 200 000	1
	172

16. Mărimea intervalelor unei distribuţii heterograde trebuie să fie fixată cu cea mai mare îngrijire. O distribuţie pe intervale mici se confundă uneori, după cum am văzut mai sus, cu o simplă înşirare de date; ea mai poate să conţină clase pentru care frecvenţa corespunzătoare este nulă. Ambele aceste aspecte trebuie evitate.

În principiu, o scară cu intervale mici scoate în relief lucruri de amănunt care nu aparţin întotdeauna caracterului regulat al fenomenului colectiv, pe când intervalele mai mari atenuează iregularităţile distribuţiei şi permit degajarea mai clară a *permanentului din accidental.* Trebuie să avem însă grijă ca prin alegerea unor intervale prea mari să nu ajungem la eclipsarea în total sau în

Scanned with OKEN Scanner

parte a trăsăturilor aceluiaşi permanent. Aceste principii mărginesc în ambele sensuri mărimea intervalelor unei serii heterograde. O analiză făcută în prealabil în lumina lor ne conduce uşor la fixarea mărimii optime a intervalelor. Însă această analiză trebuie făcută efectiv.

Este bine ca, în orice caz, chiar dacă o scară cu intervale neegale este inevitabilă, să observăm anumite reguli pentru fixarea acestora. Astfel, este de dorit ca intervalele mici să se poată deduce din cele mari prin simplă diviziune, încât, dacă avem un interval egal cu 5 000, nu vom adopta pentru mărimea intervalul inferior nici numărul 2 000, nici 3 000, ci 2 500 sau 1 000 sau 500.

Este iarăşi recomandabil ca intervalele să fie astfel ordonate încât primele să reproducă prin adunare pe cele imediat următoare. În tabelul 13 B, de pildă, primele cinci intervale contopite în unul singur ne dau un interval egal cu cel de al şaselea. Mai departe, dacă noile intervale astfel obţinute sunt grupate două câte două, obţinem intervalele 0-9999, 10000-19999, 20000-29999, egale cu intervalul 30000-39999 din tabel. Acelaşi lucru se poate observa în tabelul 11. Avantajul unei astfel de structuri a scării cu intervale neegale este lesne de înţeles: operaţia de transformare a distribuţiei într-alta, a cărei scară să aibă intervale egale, este mult uşurată.

De multe ori, din cauza faptului că elementele cuprinse în clasele marginale ale unei distribuţii au în general un caracter excepţional şi deci constituie un punct delicat al acesteia, se obişnuieşte ca intervalele claselor marginale să fie împărţite în subintervale egale cu o fracţiune simplă din mărimea comună a celorlalte. Bineînţeles că o astfel de distribuţie nu trebuie considerată în fond ca o distribuţie cu intervale neegale; subintervalele marginale trebuie contopite în intervale întregi înainte ca distribuţia să fie supusă vreunei transformări metodologice. Interesul pe care îl prezintă studiul mortalităţii infantile în special în primele zile după naştere ne obligă să facem uz de subintervale pentru statisticile relative la decese. Aceste statistici sunt distribuţii heterograde în raport cu vârsta, intervalele fiind de obicei egale cu un an. Or, pentru cunoaşterea completă a mortalităţii infantile, intervalul de un an este prea mic; iată de ce este nevoie de a se introduce subintervale egale cu o săptămână sau o lună. Tabelul 14 constituie un exemplu de astfel de distribuţie, care însă nu este cel mai tipic în această direcţie, subintervalele nefiind în acest caz egale între ele.

TABELUL 14

Morţii până la cinci ani, repartizaţi după vârstă

Etatea în momentul decesului	Numărul morţilor
0 - 7 zile	12 763
7 - 30 zile	10 767

Scanned with OKEN Scanner

(continuare tab. 14)

Etatea în momentul decesului	Numărul morților
1 - 6 luni	15 176
6 - 12 luni	10 706
1 - 2 ani	11 878
2 - 3 ani	15 932
3 - 4 ani	3 818
4 - 5 ani	2 694
	73 734

Extras din ***Mișcarea populației pe anii 1904-1905****, p. 46-47.*

17. *Scara în progresie geometrică*. Pentru punerea în evidență a anumitor aspecte ale distribuțiilor heterograde, s-au imaginat și alte scări decât cele analizate până acum. Tipul cel mai des utilizat este *scara în progresie geometrică*. Fiecare din limitele intervalelor unei astfel de scări se obține din precedenta, prin înmulțirea cu o *rație* constantă. Astfel, dacă limita primului interval este egală cu 100, iar rația cu 1,4, limitele intervalelor următoare vor fi respectiv egale cu:

$$100 \times 1{,}4 = 140$$
$$140 \times 1{,}4 = 196$$
$$196 \times 1{,}4 = 274$$

TABELUL 15 A

Satele din județul Gorj

Numărul locuitorilor	Numărul satelor
25 - 40	3
41 - 64	3
65 - 102	17
103 - 164	40
165 - 262	78
263 - 419	97
420 - 671	107
672 - 1 074	56
1 075 - 1 718	19
1 719 - 2 749	3
	433

Scanned with OKEN Scanner

Vom înțelege mai bine utilitatea adoptării unei astfel de scări după ce vom vorbi despre scara logaritmică. Deocamdată, dăm cu titlu de exemplu distribuția satelor din județul Gorj în raport cu numărul locuitorilor, distribuție făcută pe o scară cu intervalele în progresie geometrică (tabelul 15 A). Pentru posibilitatea unei înțelegeri a caracterului acestui fel de scară, dăm în același timp și distribuția acestor sate pe o scară cu intervale egale (tabelul 15 B).

Din comparația acestor două tabele se vede că, pe când în tabelul 15 B frecvențele se concentrează în primele intervale, în tabelul 15 A, din cauza concentrării frecvențelor în intervalele din mijlocul distribuției, aceasta se prezintă sub o formă oarecum simetrică. Rația progresiei geometrice formate din limitele intervalelor distribuției din tabelul 15 A este egală cu 1,6.

TABELUL 15 B

Satele din județul Gorj

Numărul locuitorilor	Numărul satelor
0 - 99	19
100 - 199	69
200 - 299	74
300 - 399	63
400 - 499	66
500 - 599	39
600 - 699	29
700 - 799	20
800 - 899	13
900 - 999	13
1 000 - 1 099	8
1 100 - 1 199	4
1 200 - 1 299	5
1 300 - 1 399	6
1 400 - 1 499	-
1 500 - 1 599	1
1 600 - 1 699	-
1 700 - 1 799	3
1 800 - 1 899	-
1 900 - 1 999	1
	433

18. *Limitele intervalelor unei scări.* O altă chestiune ridicată de clasificarea sub formă de distribuție heterogradă este aceea a modului în care se precizează limitele intervalelor ce alcătuiesc scara unei astfel de distribuții.

Scanned with OKEN Scanner

Formele cele mai des întrebuințate în acest scop sunt următoarele patru:

(a)

mai mare sau egal cu 0 și mai mic decât 1
" " " " " 1 " " " " 2
" " " " " 2 " " " " 3
" " " " " 3 " " " " 4
" " " " " 4 " " " " 5

(b)	(c)	(d)
0 - 0,9	0 - 1	0 -
1 - 1,9	1 - 2	1 -
2 - 2,9	2 - 3	2 -
3 - 3,9	3 - 4	3 -
4 - 4,9	4 - 5	4 -

Forma cea mai corectă, fiind cea mai explicită, este forma (a). Această formă evită orice ambiguitate și cuprinde toate valorile din scară, fără ca pentru aceasta să fie nevoie de vreo convenție suplimentară; are dezavantajul de a fi greoaie și nu prea sugestivă. De altfel, spațiul prea mare pe care-l ocupă a făcut ca întrebuințarea ei în publicațiile statistice să devină din ce în ce mai rară. Formele (b) și (c) sunt în această privință preferabile.

Trecând acum la analiza acestor două din urmă, trebuie să observăm, în primul rând, că forma (b) lasă impresia că nu toate valorile sunt cuprinse în scară. Astfel, valoarea 0,95 pare că nu este cuprinsă nici în primul interval, care se termină cu valoarea 0,9, nici în al doilea interval care începe de la 1,0. Această lacună este aparentă numai. Sunt două moduri de a interpreta scara (b). Se poate presupune în primul rând că toate valorile au fost exprimate la prima zecimală, așa încât valoarea 0,95, devenind, prin aproximare, egală cu 1,0, va face parte din intervalul 1-1,9. Astfel privită chestiunea, forma (b), deși permite încadrarea oricărei valori într-unul din intervale, prezintă un aspect curios: intervalele nu încep în realitate de la valorile indicate în scară. De pildă, intervalul indicat prin limitele 1,0-1,9 conține de fapt valorile mai mari sau egale cu 0,95 și mai mici decât 1,95, încât, pentru reprezentarea scării (b) prin forma (a), limitele intervalelor trebuie schimbate.

Aceste neajunsuri ale formei (b) sunt în parte atenuate dacă îi dăm o altă interpretare; primul interval va cuprinde valorile care au la partea întreagă 0, prima zecimală putând fi oricare din cifre până la 9 inclusiv, al doilea interval va cuprinde valorile ce au 1 la partea întreagă etc. Acest nou fel de a interpreta forma (b) este complet diferit de cel precedent. Dovada ne-o face faptul că valoarea 0,95, considerată mai sus, nu va mai fi cuprinsă în al doilea interval, ci în primul.

Neajunsul principal al formei (b) constă însă din faptul că intervalele par egale cu 0,9, deși, după ambele interpretări date acesteia, mărimea lor este de fapt egală cu 1. Acest neajuns este specific formei (b). Trebuie să recunoaștem

Scanned with OKEN Scanner

însă că singura formă (b) ne dă implicit o indicație asupra gradului de aproximație adoptat în măsurătorile făcute.

Forma (c) este cea mai des întrebuințată; însă ambiguitățile la care poate da naștere nu sunt evitate decât printr-o convenție suplimentară. Această convenție constă din faptul că numai limitele inferioare sunt considerate ca făcând parte din intervalul respectiv. Astfel, valoarea 1, care putea fi repartizată și în primul și în cel de-al doilea interval, în urma convenției făcute, ea aparține neîndoios intervalului 1-2. Există, de asemenea, o formă (d) și mai simplă, dedusă din forma (c) sau (b) prin suprimarea limitelor superioare ale intervalelor, care nu mai au nevoie să fie repetate, deoarece sunt egale cu limitele imediat următoare.

Fapt cert este că nici azi divergențele părerilor în privința celei mai indicate forme de adoptat pentru scara unei distribuții heterograde n-au dispărut și că o unificare în această direcție pare să întârzie. Cu atât mai mult statisticianul trebuie să fie atent la felul în care este exprimată scara.

În principiu, distribuțiile în raport cu o variabilă într-adevăr continuă se reprezintă mai comod pe o scară de tipul (c). Forma (b) este de preferat în cazul distribuțiilor de variabilă discontinuă, chiar atunci când aceasta, după cum am văzut, poate să fie asimilată unei variabile continue. Pentru astfel de distribuții, chestiunea neajunsurilor scării (b), așa cum le-am arătat mai sus, nu se mai poate pune, deoarece dispare însăși factorul de care depindea și care era continuitatea efectivă a variabilei.

Acestea sunt normele după care ne-am condus în ceea ce privește scara distribuțiilor ce se întâlnesc în aceste pagini.

19. *Distribuții închise și deschise*. Intervalele marginale ale unei distribuții heterograde joacă un rol special în statistică. Am văzut mai sus cum uneori este nevoie de a le împărți în subintervale. Alteori limitele acestor intervale nu sunt precizate. În acest caz, distribuția poartă numele de distribuție *deschisă* la unul sau ambele capete, spre deosebire de cazul când limitele acestora sunt precizate și când distribuția se numește *închisă*.

Neprecizia care decurge din forma unei distribuții deschise face ca o astfel de distribuție să fie foarte greu utilizabilă. În orice caz, prudența nu poate fi niciodată prea mare atunci când este vorba de a utiliza o distribuție deschisă. Iată de ce prezentarea datelor sub forma de distribuție închisă este un deziderat important al metodei statistice. Totuși, acest deziderat nu poate fi întotdeauna satisfăcut. Acesta este cazul ultimului interval al distribuției din tabelul 13 B, care a fost lăsat înadins deschis pentru ca să cuprindem în distribuție și cazul municipiului București. Neajunsul unui interval deschis este mai accentuat în cazul când acesta se găsește la marginea superioară a scării. Majoritatea distribuțiilor deschise sunt din nefericire deschise la capătul acesta. O distribuție deschisă poate să fie făcută utilizabilă dacă i se adaugă un supliment de informații relative la intervalele deschise. Astfel de informații pot fi mediile claselor alcătuite de aceste intervale sau valorile extreme ce sunt încadrate în distribuție. În cazul tabelului 13 B, valorile extreme sunt cele date mai sus: 871 pentru marginea inferioară și 631 288 pentru cea superioară.

Scanned with OKEN Scanner

20. *Distribuţiile continue şi discontinue.* Distribuţiile heterograde se mai pot deosebi şi din punct de vedere al naturii variabilei independente; sunt distribuţii discrete sau discontinue şi distribuţii continue.

Conform analizei variabilelor întâmplătoare pe care am făcut-o într-unul din paragrafele precedente, fac parte din categoria distribuţiilor continue distribuţiile din tabelele 12, 13 A ,13 B, 14.

Tabelul 16 ne înfăţişează o distribuţie discontinuă.

Vom vedea mai târziu că diferitele prelucrări ale distribuţiilor vor trebui să fie conduse în mod diferit, după cum vom avea de-a face cu o distribuţie continuă sau discontinuă.

TABELUL 16

Familiile din Franţa, repartizate după numărul copiilor în viaţă

Numărul copiilor în viaţă	Numărul familiilor
0	1 881 264
1	3 674 348
2	2 809 130
3	1 525 551
4	816 542
5	449 313
6	255 827
7	142 524
Peste 7	151 750
	11 706 249

Extras din ***Statistique des familles en 1926*** *(publicaţia Statisticii Generale a Franţei), p.70.*

21. *Frecvenţe relative.* Elementul prin care o distribuţie homo sau heterogradă devine un instrument util de cercetare ştiinţifică este în primul rând *frecvenţa.* Frecvenţa - care nu este altceva decât numărul care ne arată câte cazuri din cele observate au corespuns unei clase a scării de distribuţie - poate fi *absolută* sau *relativă.* Dacă într-o comună s-au născut în cursul unui an 4 362 băieţi şi 4 123 fete, frecvenţa absolută a naşterilor masculine este chiar numărul 4 362, fără nici o altă adăugire. Prin raportarea într-un mod oarecare a frecvenţei absolute (4 362) la totalul naşterilor (8 485), obţinem ceea ce numim frecvenţă relativă. Frecvenţa relativă este aceea care ne dă o idee exactă de importanţa clasei corespunzătoare faţă de ansamblul de cazuri observate. A spune că s-au născut 4 362 de băieţi e totuna cu a nu spune nimic; trebuie să precizăm anume din ce total de naşteri. De fapt, raportarea frecvenţelor absolute la totalul lor este o operaţie pe care o facem chiar fără să trecem la calculul efectiv al frecvenţelor relative. Orice serie statistică este constituită

Scanned with OKEN Scanner

dintr-un ansamblu de date care pune imediat în evidență variația frecvențelor de la un caz la altul, calculul nu face decât să precizeze gradul de variație. Numai dacă ar exista serii statistice formate dintr-o singură dată am putea spune că frecvența absolută există într-adevăr. Această distincție între modurile cum este reprezentată o frecvență are totuși utilitatea ei metodologică.

Raportarea frecvenței absolute la totalul frecvențelor se face de obicei prin proporția la 1, 10, 100, 1000 etc.

Astfel, dacă n este frecvența absolută, N, totalul frecvențelor din distribuție, frecvența relativă f va fi:

$$f = \frac{n}{N}, \frac{n \times 10}{N}, \frac{n \times 100}{N}, \ldots\ldots\ldots$$

după cum aceasta va fi raportată la unitate, la 10, la 100 etc. În exemplul citat mai sus avem frecvențele la sută:

naşteri masculine	51,4
naşteri feminine	48,6

Sunt câteva lucruri de observat în legătură cu frecvențele relative. În primul rând, numerele ce le reprezintă pot fi fracționare, spre deosebire de frecvențele absolute care nu pot fi decât întregi. Frecvențele raportate la unitate, la 100 etc. se pot deduce unele din altele printr-o simplă permutare convenabilă a virgulei zecimale. În fine, trebuie să remarcăm că suma frecvențelor raportate la unitate, 10, 100 etc. este respectiv egală cu 1, 10, 100...

Tabelul 5 constituie un exemplu de serie statistică în care frecvențele absolute sunt puse față în față cu frecvențele relative corespunzătoare.

22. *Distribuții combinate*. Pentru clasificarea într-o serie statistică a elementelor observate, este neapărat nevoie de a se face abstracție de orice atribut al acestora, altul decât cel în raport cu care se face clasificarea. Într-o clasificare a unor indivizi după înălțime nu vom ține seama de faptul că aceștia pot fi de greutăți și de vârste diferite, că nu au aceeași culoare de ochi etc. Dacă e vorba despre o serie în timp privind producția agricolă a unei țări, toate boabele ce alcătuiesc recolta aceluiași an trebuie considerate identice, indiferent de calitatea lor, de regiunea de unde provin.

Se poate întâmpla însă ca să fie necesar să se urmărească manifestările unui fenomen, dintr-o dată, din două sau mai multe puncte de vedere. În acest caz, pentru prezentarea datelor statistice, se utilizează forme noi, obținute prin combinări între cele trei serii tipice studiate până acum. Aceste forme de prezentare a datelor statistice sunt foarte variate și nu mai pot fi încadrate în grupe caracteristice. Tabelul 17 face parte din această categorie; cazurile examinate au fost clasificate din trei puncte de vedere: al neamului, al vârstei și al sexului. Dintr-un astfel de tabel se pot extrage serii statistice în raport cu un singur atribut. Astfel, din tabelul 17 se deduc tabelele 17 A, 17 B, 17 C.

Scanned with OKEN Scanner

Este de cea mai mare importanţă să se observe că o astfel de operaţie nu este reversibilă: plecând de la tabelele 17 A, 17 B, 17 C, nu se poate reconstrui tabelul 17, ceea ce înseamnă că a urmări manifestările unui fenomen din mai multe puncte de vedere *dintr-o dată* şi *separat* sunt două lucruri complet deosebite. Prima operaţie este mult mai largă şi cuprinde pe cea de-a doua. Acest fapt trebuie să fie pătruns de cei ce se ocupă cu culegerea datelor, precum şi de cei ce vor să le analizeze numai.

TABELUL 17

Cetăţenii români emigraţi în cursul anului 1930

Neamul	Sub 15 ani		Între 15-25 ani		Între 25-55 ani		Peste 55 ani		Total
	B.	F.	B.	F.	B.	F.	B.	F.	
Români	32	40	31	66	262	163	6	8	608
Unguri	79	85	86	128	558	286	11	7	1 240
Ucranieni	28	34	97	102	608	93	13	12	987
Evrei	428	468	394	548	848	810	115	92	3 703
Germani	371	412	362	475	987	693	35	27	3 362
Polonezi	6	3	6	7	24	7	-	-	53
Ruşi	31	24	24	26	151	47	3	5	311
Iugoslavi	20	29	33	28	137	68	1	-	316
Bulgari	52	58	11	11	58	48	7	2	247
Cehoslovaci	-	3	4	1	4	1	-	-	13
Alte neamuri	4	4	2	-	8	3	-	-	21
Total	**1 051**	**1 160**	**1 050**	**1 392**	**3 645**	**2 219**	**191**	**153**	**10 861**

*Extras din **Anuarul statistic al României pe anul 1930**, p.242.*

TABELUL 17 A

Vârsta	Numărul emigranţilor
Sub 15 ani	2 211
Între 15 şi 25 ani	2 442
Între 25 şi 55 ani	5 864
Peste 55 ani	344
Total	**10 861**

TABELUL 17 B

Sexul	Numărul emigranţilor
Bărbaţi	5 937
Femei	4 924
Total	**10 861**

Scanned with OKEN Scanner

TABELUL 17 C

Neamul	Numărul emigranților
Evrei	3 703
Germani	3 362
Unguri	1 240
Ucrainieni	987
Români	608
Iugoslavi	316
Ruși	311
Bulgari	247
Polonezi	53
Cehoslovaci	13
Alte neamuri	21
Total	**10 861**

Scanned with OKEN Scanner

Capitolul III

Reprezentarea grafică

23. *Generalități.* Caracterele manifestărilor fenomenelor colective, așa cum reies din ansamblul de date statistice prezentate sub formă de serii, sunt susceptibile de cele mai multe ori de a fi scoase în relief prin reprezentarea grafică a acestor serii. Utilitatea reprezentării grafice este o consecință a faptului că, dintre toate organele noastre de simț ochiul pare să fie cel mai bine organizat pentru a aprecia evenimentele exterioare. Ochiul poate face imediat o sinteză a impresiilor din afară, apreciind: 1) forma, 2) direcțiile, 3) lungimile și mai ales raportul dintre acestea din urmă.

Iată de ce o reprezentare grafică poate să scoată în evidență lucruri pe care șirul de date le conține de fapt, dar care ar fi mai greu de descoperit altfel.

Comparația manifestărilor aceluiași fenomen în condiții diferite - actul științific cel mai important - este mult ușurată de reprezentarea grafică: ochiul compară mai ușor două grafice decât mintea, două șiruri de numere.

Reprezentarea grafică ne mai poate fi utilă și din alt punct de vedere. Ea permite uneori găsirea în mod grafic a unor valori care ar necesita operații mai complicate dacă ar fi calculate direct. Bineînțeles, nu este vorba aici decât de unele valori aproximative, calea grafică neputând fi adoptată decât atunci când o precizie mai mare nu este absolut necesară.

În fine, pentru o expunere mai clară a metodelor de lucru, reprezentarea grafică este un auxiliar de primul ordin. Acesta este, în definitiv, motivul pentru care, înainte chiar de a intra în materia propriu-zisă a metodologiei, vom expune în paragrafele următoare principiile esențiale ale acestei părți a metodei statistice, urmând ca, pe măsură ce expunerea elementelor analitice necesare ne-o vor permite, să ne ocupăm și de chestiunile de un ordin mai special.

24. *Diagrame.* Vom începe cu reprezentarea grafică a distribuțiilor homograde, pentru că modul de reprezentare a acestora constituie întrucâtva esența reprezentării grafice: redarea prin desen a rapoartelor numerice. Graficul corespunzător poartă numele de diagramă, deși acest termen este întrebuințat deseori într-un sens mai larg, pentru denumirea oricărui grafic; sensul restrâns al lui este cel la care ne oprim în statistică.

Scanned with OKEN Scanner

TABELUL 18

Bolnavii operaţi în spitalele eforiei, în cursul anului 1930

Nr.de ordine	Felul bolilor	Numărul bolnavilor	La suta de bolnavi
I	Chirurgie	2 410	41
II	Obstetrică şi ginecologie	1 186	20
III	Oto-rino-laringologie	774	13
IV	Căi urinare	488	8
V	Oftalmologie	338	6
VI	Boli venerice	314	5
VII	Ginecologie	201	4
VIII	Boli interne	161	3
		5 872	**100**

*Extras din **Anuarul statistic al României, 1930**, p.363.*

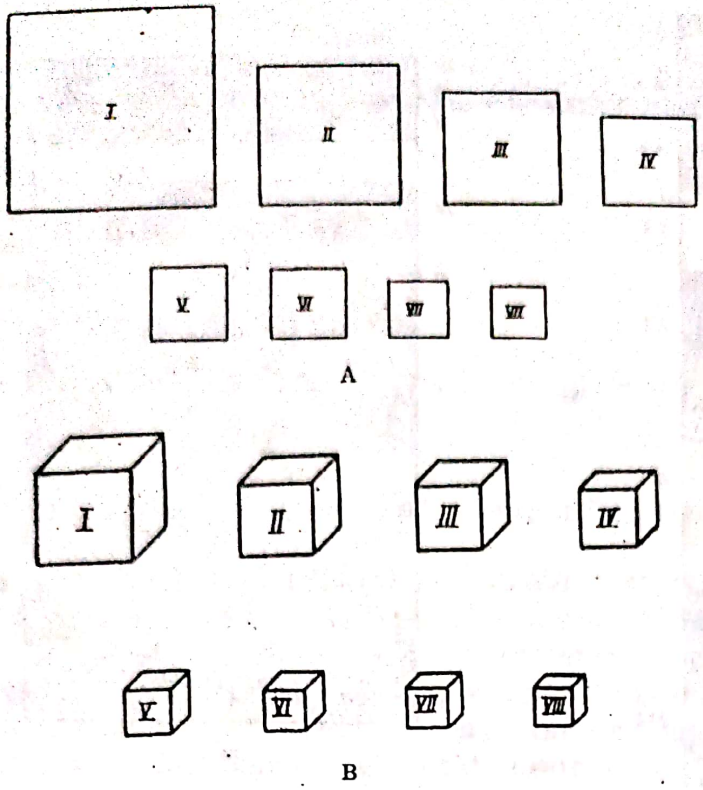
Diagrama acestui tabel poate să ia formele cele mai diverse; le vom menţiona pe toate, analizând în acelaşi timp avantajele şi neajunsurile pe care le reprezintă fiecare.

a) Numărul de bolnavi din diferite clase se reprezintă prin pătrate, cercuri sau cuburi, ale căror suprafeţe, respectiv volume, să fie *pe grafic* proporţionale cu frecvenţele corespunzătoare (graficul 1). Pentru construcţia unei astfel de diagrame trebuie extrase rădăcinile pătrate ale frecvenţelor. Laturile pătratelor, razele cercurilor vor fi proporţionale cu aceste rădăcini. În cazul reprezentării prin cuburi, laturile acestora vor fi proporţionale cu rădăcina cubică a frecvenţelor.

Acest mod de a reprezenta grafic o distribuţie are un inconvenient care face ca avantajul cel mai important al reprezentării grafice să fie întrucâtva întunecat: raportul numeric dintre frecvenţe nu mai reiese aşa de intuitiv, ochiul apreciind în primul rând lungimile şi mai puţin suprafeţele şi volumele. Totuşi el este cel mai utilizat atunci când frecvenţele diferitelor clase trebuie repartizate pe grafic în diferite puncte, cum ar fi cazul unei distribuţii în spaţiu.

b) Inconvenientele diagramei de tipul precedent sunt în parte înlăturate, dacă se adoptă pentru reprezentarea grafică a frecvenţelor un dreptunghi împărţit în mai multe altele. În acest caz, singură înălţimea dreptunghiurilor divizionare este cea care variază proporţional cu mărimea frecvenţei. Fiind vorba de o singură dimensiune, ochiul poate aprecia mai just raporturile dintre înălţimi. Graficul 2 ne arată care pot fi în acest caz diagramele corespunzătoare distribuţiei din tabelul 18.

Scanned with OKEN Scanner



Tabelul 18

Graficul 1. Reprezentarea grafică a unei distribuţii homograde.

Diferenţa între impresiile pe care ni le fac diagramele A şi B ne arată că şi acest mijloc de reprezentare grafică nu este lipsit de inconveniente; o bază mare şi o înălţime mică tind să exagereze importanţa numerică a frecvenţelor mari în detrimentul celor mici.

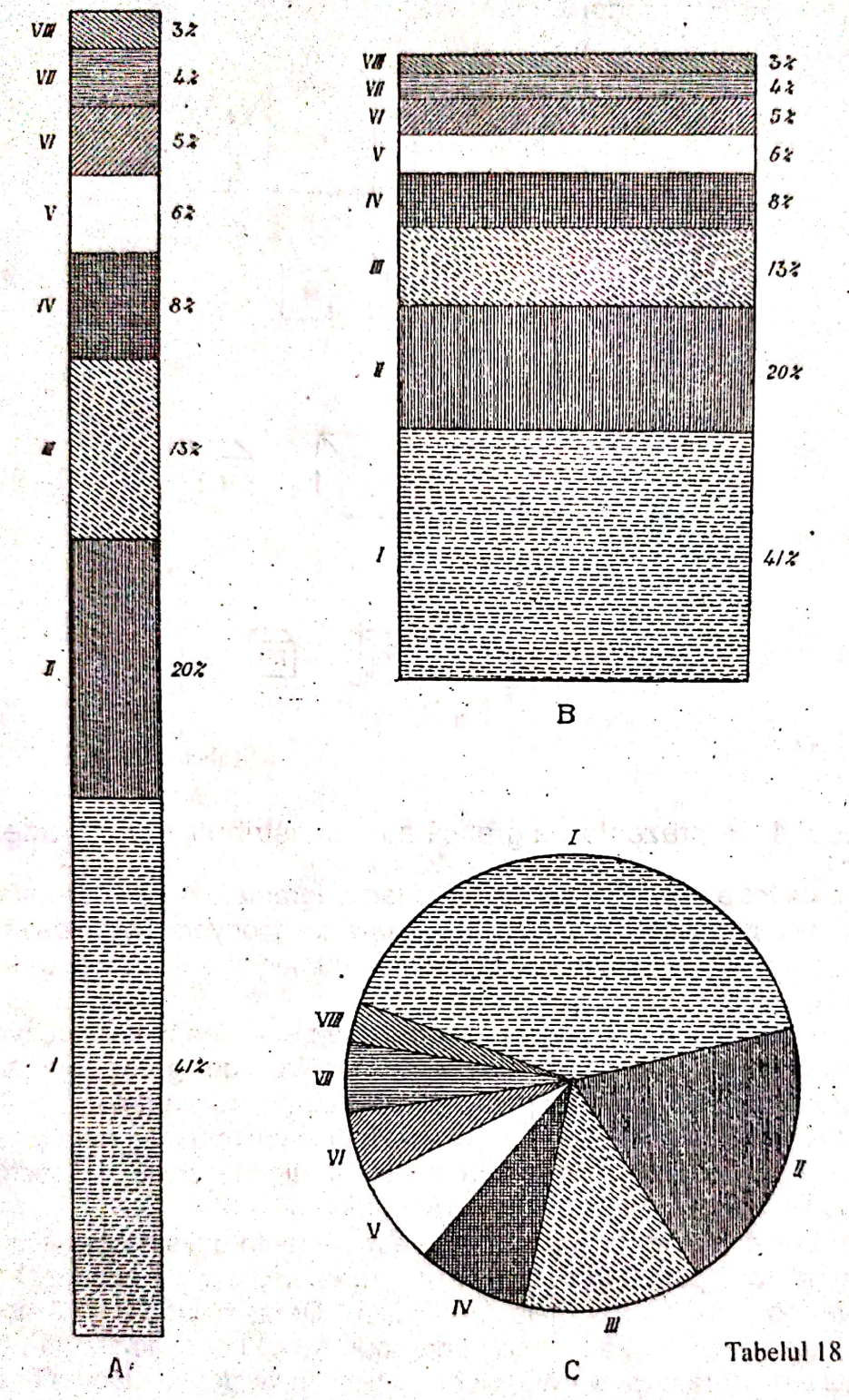
În aceeaşi categorie de diagrame vom pune pe acelea în care frecvenţele sunt reprezentate prin sectoare de cerc, numărul de grade ale arcelor corespunzătoare acestora fiind proporţional cu mărimea frecvenţelor[1].

Acest tip de diagramă, deşi nu mai prezintă inconvenientul menţionat mai sus, nu este prea sugestiv, din cauza direcţiilor diferite ce le au sectoarele reprezentând frecvenţele (graficul 2, diagrama C).

c) Mai există, în fine, încă un mod de a reprezenta o distribuţie: acela prin care se face să corespundă frecvenţelor nişte dreptunghiuri de aceeaşi bază şi de înălţimi proporţionale cu primele (graficul 3). Dreptunghiurile sunt aşezate astfel ca una din bazele lor să fie în linie dreaptă. Acest tip de diagramă permite să i se alăture o scară care să pună şi mai bine în evidenţă raporturile dintre frecvenţe.

---

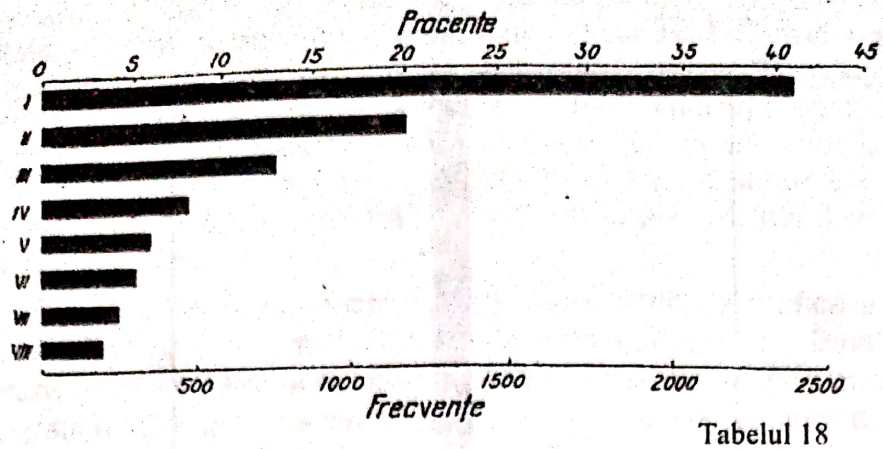
[1] *Se obişnuieşte ca să se lase neutilizat un cerc concentric cu cel ce alcătuieşte diagrama; aceasta, pentru a evita regiunea neclară care rezultă în desen din concurenţa în centrul cercului a diferitelor raze ale acestuia.*



Tabelul 18

Graficul 2. Diagrame dreptunghiulare și în cerc pentru reprezentarea seriilor homograde.

Scanned with OKEN Scanner



Tabelul 18

Graficul 3. Reprezentarea grafică prin bare a distribuțiilor homograde.

Personal, înclin să cred că așezarea orizontală a dreptunghiurilor unei astfel de diagrame este mai favorabilă scopurilor urmărite de reprezentarea grafică[1]. Așezarea verticală a lor este totuși de preferat când între diferitele clase ale distribuției se poate stabili o ordine graduală. Se știe că în acest caz seria statistică este prezentată în această ordine, ordine care trebuie să fie păstrată și în diagramă. Un astfel de grafic permite urmărirea variației frecvențelor în raport cu gradul atributului distribuției. Acesta este cazul tabelului 19, căruia îi corespunde graficul 4.

TABELUL 19

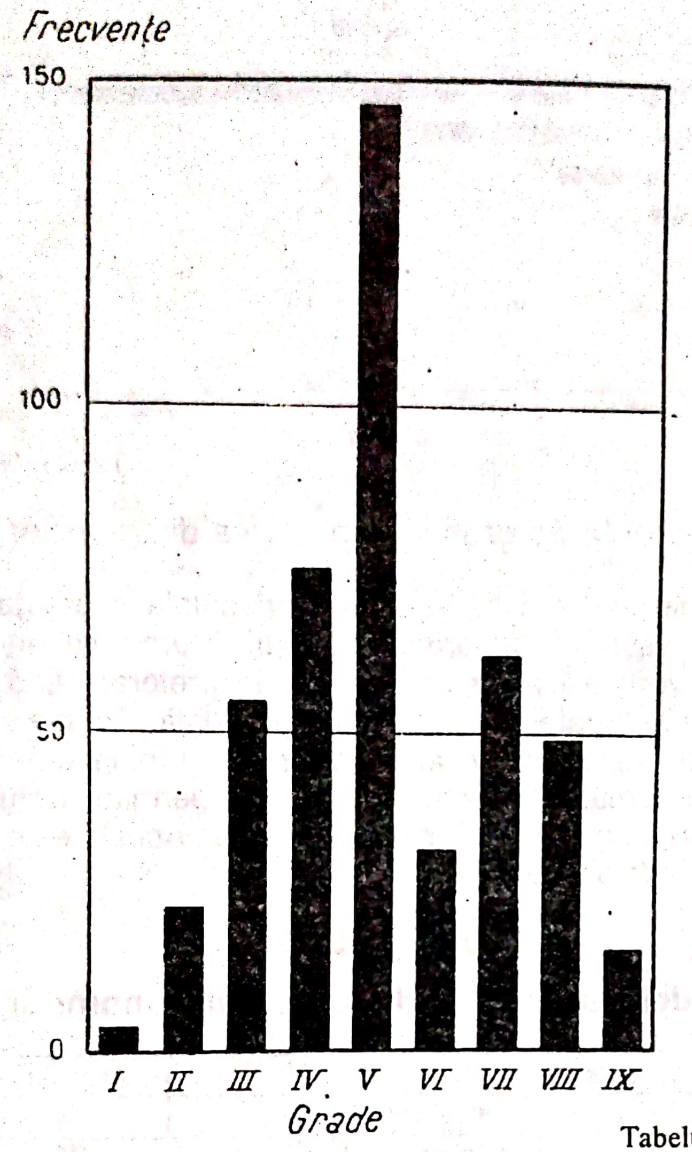
Gradele inginerilor din corpul agronomilor

Nr. de ordine	Gradul	Ingineri
1	Ingineri stagiari	4
2	Ingineri cl. II	23
3	Ingineri cl. I	53
4	Ingineri șefi cl. II	75
5	Ingineri șefi cl. I	146
6	Inspectori cl. II	32
7	Inspectori cl. I	62
8	Inspectori generali	59
9	Consilieri	16
		740

Reprodus din ***Statistica funcționarilor statului, 1932****, p. 42.*

[1] *N-am întâlnit nicăieri această afirmație, care totuși nu mi se pare lipsită de temei. Faptul poate să fie o consecință a poziției orizontale a liniei ochilor. De altfel, oricine îl poate controla: un segment de linie dreaptă se împarte din ochi în mai multe părți egale mai repede și cu o mai bună aproximație dacă este așezat paralel cu linia ochilor.*

Scanned with OKEN Scanner



Tabelul 19

Graficul 4. Reprezentarea grafică a unei distribuţii homograde în raport cu un atribut gradual.

Aceste dreptunghiuri nu trebuie lipite unul de altul, ci aşezate în mod echidistant. Unii statisticieni, mai ales atunci când distribuţia nu cuprinde decât 2-3 clase, aşază dreptunghiurile mai mici în interiorul celor mari sau *à cheval* pe marginea acestora. Este un procedeu care poate fi considerat periculos, deoarece această încălcare de suprafeţe poate falsifica aprecierea raportului dintre mărimile lor.

Este iarăşi recomandabil ca, în general, să fie înscrise pe diagramă în dreptul fiecărei suprafeţe şi frecvenţele relative corespunzătoare. Când ochiul apreciază un fapt, prezenţa frecvenţelor relative permite un control imediat şi o precizare a acestei aprecieri.

Scanned with OKEN Scanner

Am analizat până aici tipurile de diagrame cele mai des întrebuințate; multe altele au fost și mai pot fi imaginate. O singură normă trebuie să ne călăuzească în construirea diagramelor: simplitate, claritate, renunțare la tot ce poate fi artificial. Să nu se creadă, de exemplu, că o diagramă ce utilizează figuri pentru reprezentarea distribuției este mai sugestivă decât vreuna din tipurile descrise. Principiul reprezentării unei distribuții prin diagramă este așa de simplu, încât ori cel ce o are în față o poate pricepe, ori, dacă nu, oricare alt artificiu este de prisos.

25. *Cartograme*. În ceea ce privește reprezentarea grafică a distribuțiilor în spațiu, au fost și sunt adoptate formele cele mai variate. Bineînțeles că o reprezentare grafică a unei astfel de distribuții trebuie să se facă cu ajutorul unei hărți geografice. De aici, termenul de *cartogramă* ce se întrebuințează, de obicei, pentru denumirea graficelor din această categorie.

TABELUL 20

Importul României, pe țări

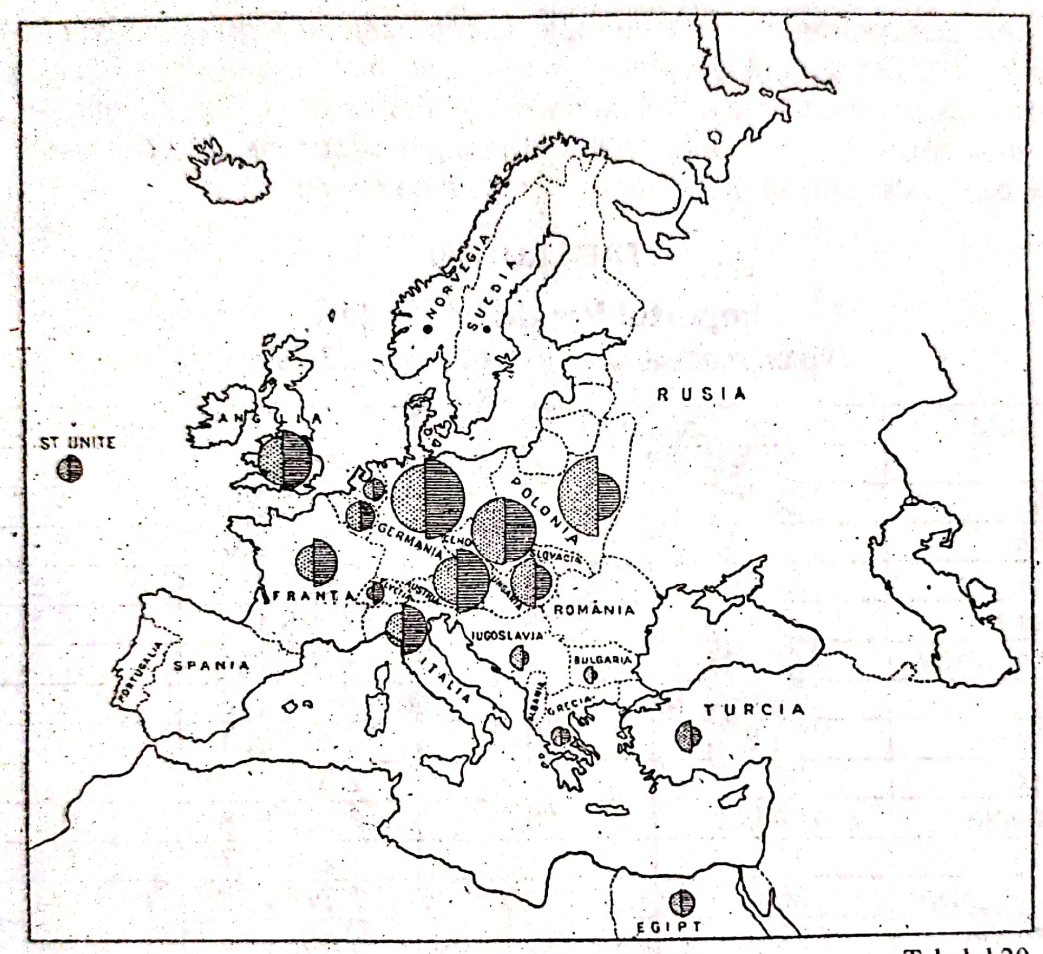
(Valori medii anuale pentru anii 1923-1927)

Țara	Cantitatea		Valoarea	
	Tone	%	Mii lei	%
Anglia	58 274	6,7	2 696 074	9,2
Austria	71 712	8,2	4 390 982	15,0
Belgia	22 289	2,6	672 853	2,3
Bulgaria	7 755	0,9	25 131	0,1
Cehoslovacia	140 103	16,1	3 761 431	12,8
Egipt	15 680	1,8	581 278	2,0
Elveția	3 926	0,4	479 961	1,6
Franța	32 137	3,7	2 332 127	7,9
Germania	134 080	15,4	6 151 078	21,0
Grecia	11 184	1,3	254 807	0,9
Italia	35 025	4,0	2 774 224	9,4
Iugoslavia	12 486	1,4	80 476	0,3
Norvegia	1 218	0,1	25 050	0,1
Olanda	8 662	1,0	420 154	1,4
Polonia	201 306	23,1	2 358 818	8,0
Spania	394	0,0	24 457	0,1
Statele-Unite	11 364	1,3	611 622	2,1
Suedia	572	0,1	53 619	0,2
Turcia	26 147	3,0	278 404	0,9
Ungaria	70 238	8,1	1 152 761	3,9
Alte țări	6 908	0,8	222 854	0,8
Total	**871 460**	**100,0**	**29 348 161**	**100,0**

Tabel calculat după datele din ***Anuarul statistic al României****, p. 274-77.*

Scanned with OKEN Scanner

Pe hartă datele statistice sunt reprezentate, în general, prin suprafețe proporționale cu ele. Din cauză că imaginile grafice ale acestor date trebuie să fie răspândite pe grafic în diferite puncte ce depind de distribuția în spațiu corespunzătoare, modul cel mai des întrebuințat pentru reprezentarea lor este acela în care aceste imagini sunt constituite din cercuri. Pentru construcția acestor grafice trebuie să ținem seamă, în linii generale, de aceleași norme după care ne conducem pentru formarea diagramelor.



Tabelul 20

Graficul 5. Importul României, pe țări de origine.

O simplă ilustrare a unuia din modurile de reprezentare a unei distribuții în spațiu este reprezentarea grafică a importului României în raport cu țara de origine (graficul 5). Graficul este construit cu ajutorul tabelului 20. Semicercurile punctate reprezintă cantitatea importată din țara corespunzătoare, raportată la cantitatea totală a importului; aceeași semnificație, însă în raport cu valoarea importurilor, o au semicercurile haşurate. Graficul acesta învederează într-un chip special utilitatea reprezentării grafice. Dintr-o singură privire se degajează din el țările din care noi importăm mai mult articole fabricate și pentru care deci semicercul haşurat este mai mare decât cel punctat (este știut că prețul mediu

Scanned with OKEN Scanner

pe cantitate este mai mic pentru produsele brute ale solului sau subsolului decât pentru articolele fabricate).

26. *Historiograme*. Reprezentarea grafică a unei serii în timp este bazată pe un principiu introdus de DESCARTES şi cunoscut în matematici sub numele de principiul reprezentării prin *coordonate carteziene*. Spuneam că seriile statistice ni se înfăţişează ca o corespondenţă între două şiruri de valori. Pentru ca această corespondenţă să fie evidenţiată în mod grafic, se consideră două drepte perpendiculare între ele şi pe care se stabileşte în prealabil un sens. Aceste drepte poartă numele de axe de coordonate. Pe grafic, una din ele este de obicei orizontală şi se numeşte axa absciselor, cealaltă, care natural va fi verticală, poartă numele de axa ordonatelor. Sensul obişnuit este de la stânga spre dreapta pentru axa absciselor, de jos în sus pentru axa ordonatelor. Pe prima se însemnează valorile variabilei independente, pe cea de-a doua, valorile funcţiei. Fiecărei perechi de valori a unei serii statistice îi va corespunde în planul axelor un punct: intersecţia dreptelor ridicate perpendicular pe axele de coordonate în dreptul valorilor corespunzătoare.

TABELUL 21

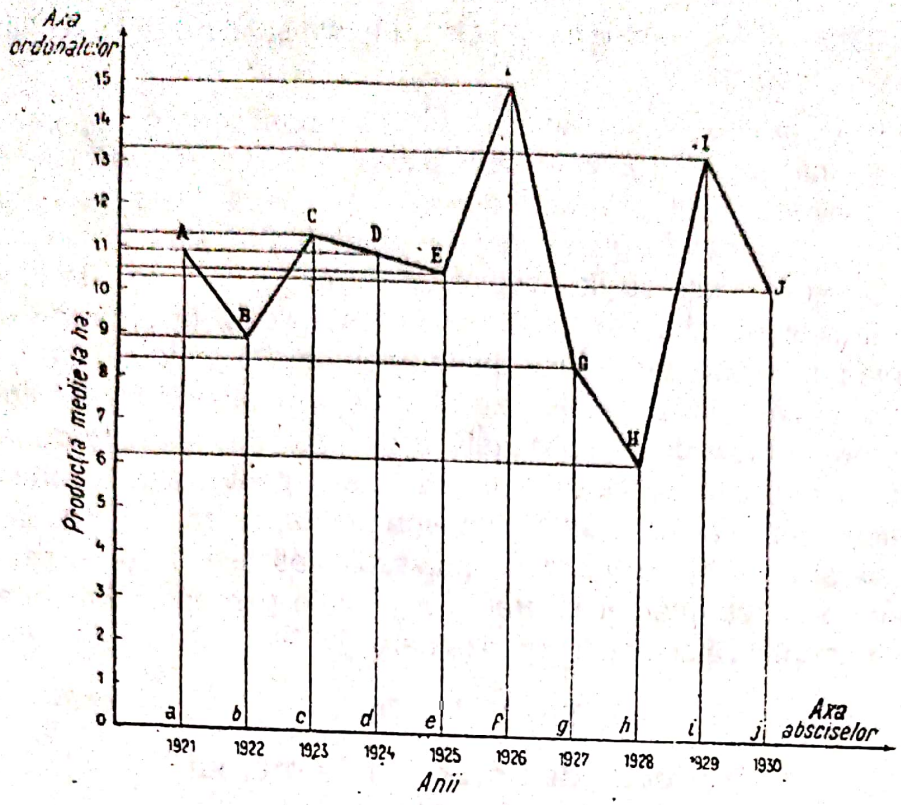
Producţia de porumb a României (1921-1930)

Anii	Producţia medie la ha în chintale
1921	10,9
1922	8,9
1923	11,3
1924	10,9
1925	10,5
1926	14,9
1927	8,4
1928	6,2
1929	13,3
1930	10,3

Extras din diferite ***Anuare statistice ale României.***

Astfel, dacă avem în vedere tabelul 21, vom însemna pe una din axele graficului 6 valorile 1921,1922,..., 1930, pe alta, o scară ce ne va servi pentru reprezentarea producţiei medii la hectar. Celor 10 perechi de valori ale seriei le corespund pe grafic 10 puncte *A, B, C,..., I, J*. Distanţele *Aa, Bb,..., Ii, Jj* reprezintă producţia medie la hectar a anilor respectivi.

Scanned with OKEN Scanner



Tabelul 21

Graficul 6. Reprezentarea prin coordonate carteziene a seriilor în timp.

Este bine să ne dăm seama că, dacă punctul unde axele se întretaie poate fi considerat ca reprezentând originea valorilor producției medii, el poate să nu aibă întotdeauna aceeași semnificație pentru scara valorilor timpului. Ne-ar trebui, în acest caz, metri de hârtie pentru reprezentarea unei serii, care s-ar putea face, cu toate acestea, pe o suprafață de câțiva centimetri pătrați. Faptul că originea timpului nu figurează pe grafic nu constituie un inconvenient, deoarece timpul este o variabilă esențial relativă, așa că orice epocă poate fi adoptată pentru originea lui. Pe grafic, ca și în exprimarea numerică a evoluției unui fenomen în timp, nu ne interesează decât mărimea intervalelor de timp ce despart observațiile între ele, și nu a acelora ce s-au scurs de la epoca de origine și până în momentul observației.

O dată punctele *A, B,..., I, J* obținute, rămâne să se pună în evidență, printr-un mod oarecare, variația în timp a fenomenului. Acest lucru se face în două moduri: se formează pe segmentele *Aa, Bb,...* niște bare de grosime convenabilă sau se unesc prin linii drepte punctele *A, B,...* între ele. Oricare ar fi procedeul adoptat, graficul corespunzător unei serii în timp poartă numele de historiogramă.

27. *Forma historiogramelor.* Procedeul de a întrebuința pentru reprezentarea unei serii în timp bare de lungimi proporționale cu datele ce o constituie este recomandabil în două cazuri:

1. Când seria în timp este imaginea manifestărilor unui fenomen care nu se produce în mod continuu în timp, ci are loc o singură dată în fiecare interval de timp al seriei. În acest caz nu ar avea sens să reprezentăm printr-o linie neîntreruptă manifestările izolate ale fenomenului considerat. Acesta este cazul producției agricole.

TABELUL 22

Balanța comercială a României pe anii 1901-1915

Anii	Import (valori mii lei)	Export (valori mii lei)	Diferențe (valori mii lei)
1901	292 436	353 831	+61 395
1902	283 345	374 819	+91 474
1903	269 924	355 630	+85 706
1904	311 372	261 872	-49 500
1905	337 538	457 101	+119 563
1906	422 114	491 360	+69 246
1907	430 509	554 019	+123 510
1908	414 058	379 431	-34 627
1909	368 300	465 057	+96 757
1910	409 716	616 505	+206 789
1911	569 745	691 720	+121 975
1912	637 906	642 104	+4 198
1913	590 013	670 705	+80 692
1914	504 241	451 891	-52 350
1915	332 942	570 182	+237 240

*Extras din **Anuarul statistic al României pe anul 1930**, p.270.*

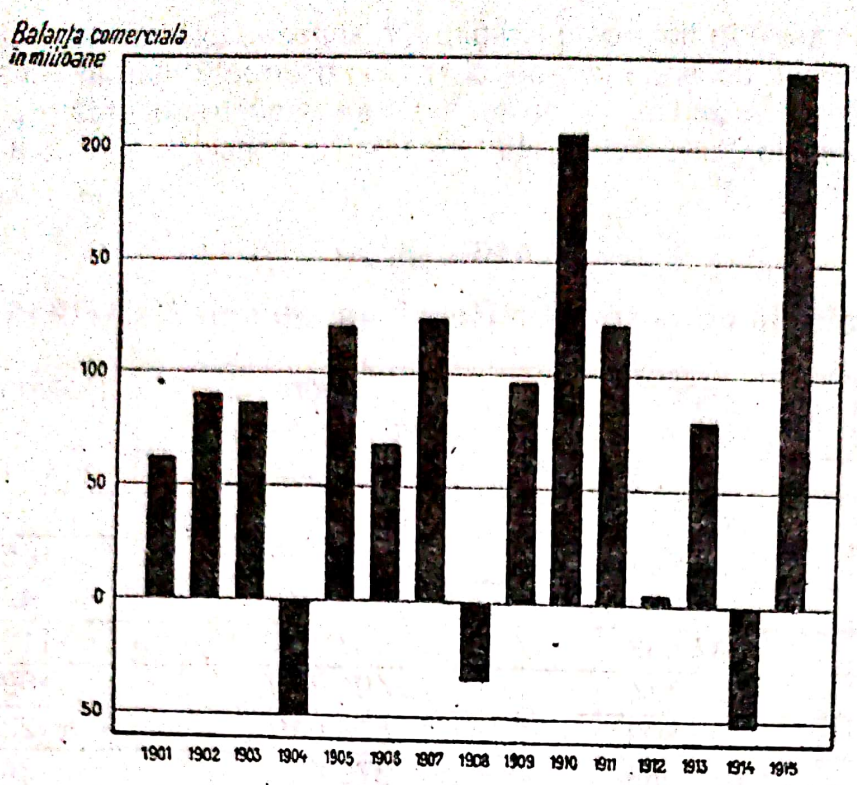
2. Când seria în timp este constituită din valori rezultate dintr-o operație de compensare. Tabelul 22 ilustrează această din urmă posibilitate. Reprezentarea grafică corespunzătoare este cea înfățișată de graficul 7.

Trebuie observat că historiograma în care se utilizează bare pentru reprezentarea grafică a datelor se apropie ca formă de diagramă.

În toate celelalte cazuri historiograma prin linia frântă este cea adoptată.

În ceea ce privește acest tip de historiograme, avem de relevat un fapt care, de cele mai multe ori, este ignorat sau desconsiderat. Să presupunem, pentru a ne fixa ideile, că avem de-a face cu o serie în timp alcătuită din date anuale. Aceste date anuale pot fi de două feluri: 1) globale, 2) medii trimestriale, lunare etc. Ultimele se obțin din cele dintâi prin împărțirea cu 4,12 etc. Există o diferență între modul cum este reprezentată o serie în timp formată de date globale și cel în care se face reprezentarea unei serii din date medii. Să considerăm tabelul 23, care cuprinde activitatea minelor de cărbuni exprimată prin cifre anuale globale și medii lunare.

Scanned with OKEN Scanner



Tabelul 22

Graficul 7. Reprezentarea grafică prin bare a unei serii în timp.

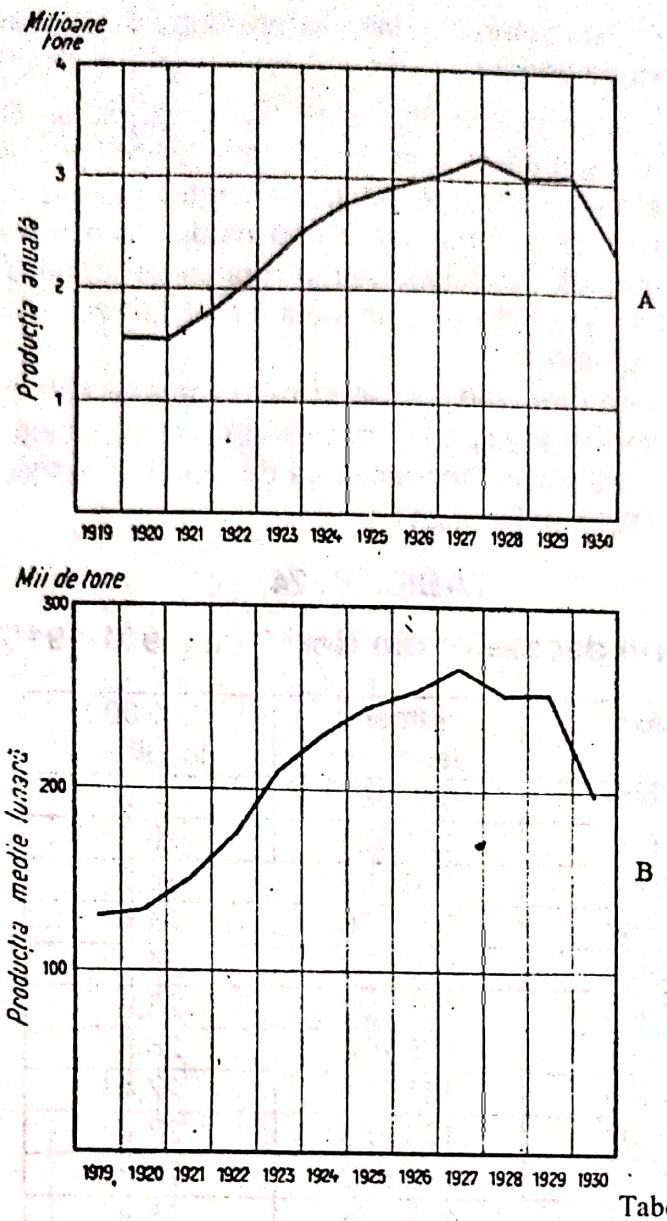
TABELUL 23

Producția de cărbuni a României pentru anii 1919-1930

Anii	Producția în tone	
	anuală	media lunară
1919	1 559 330	129 944
1920	1 587 575	132 298
1921	1 804 687	150 391
1922	2 116 221	176 352
1923	2 521 393	210 116
1924	2 776 371	231 364
1925	2 928 850	244 071
1926	3 053 553	254 463
1927	3 223 468	268 622
1928	3 027 240	252 270
1929	3 046 027	253 836
1930	2 369 882	197 490

Extras din ***Statistica minieră a României pe anul 1930****, p.131.*

Scanned with OKEN Scanner



Tabelul 23

Graficul 8. Reprezentarea grafică a datelor anuale şi a mediilor lunare ale unei serii în timp.

Valorile ce exprimă producţia globală anuală fiind rezultatul unei activităţi ce se încheie la sfârşitul anului, ordonatele ce le reprezintă trebuie să fie trase pe grafic în dreptul punctului care indică sfârşitul unui an şi începutul celui următor, aşa cum se vede în historiograma (A) a graficului 8. Dacă acum am vrea să reprezentăm evoluţia aceluiaşi fenomen, servindu-ne de mediile lunare, atunci ordonatele reprezentative trebuie aşezate în mijlocul segmentului care corespunde intervalului de timp de un an (historiograma B). Justificarea acestui principiu se găseşte în faptul că valoarea medie nu poate fi reprezentată decât

acolo unde caracterul de simetrie cuprins în sfera noţiunii acesteia n-ar avea de suferit din nici un punct de vedere[1].

28. *Principii generale pentru alcătuirea historiogramelor.* Chestiunea cea mai importantă însă în legătură cu construirea unei historiograme este aceea a raportului dintre unitatea de lungime adoptată pentru măsura timpului şi cea care serveşte pentru scara ordonatelor. După modul cum aceste unităţi sunt alese, putem avea impresia că historiograma reprezintă un fenomen cu variaţii pronunţate într-un sens şi altul sau, din contră, un fenomen stabil, ale cărui variaţii de-abia pot fi degajate.

Pentru a pătrunde mai bine modul în care raportul dintre unităţile alese pentru construcţia historiogramei poate modifica impresia vizuală pe care o face aceasta, să considerăm seria în timp alcătuită de numărul morţilor din România pentru perioada 1901-1915 (tabelul 24).

TABELUL 24

Numărul deceselor din România (1901-1915)

Anii	Numărul deceselor	La 1 000 locuitori
1901	160 361	26,5
1902	171 828	28,1
1903	156 006	25,2
1904	155 936	24,4
1905	160 173	24,7
1906	157 204	23,9
1907	175 794	26,3
1908	185 393	27,4
1909	188 325	27,4
1910	172 842	24,8
1911	179 076	25,3
1912	165 616	22,9
1913	191 689	26,1
1914	182 949	23,5
1915	193 574	24,5

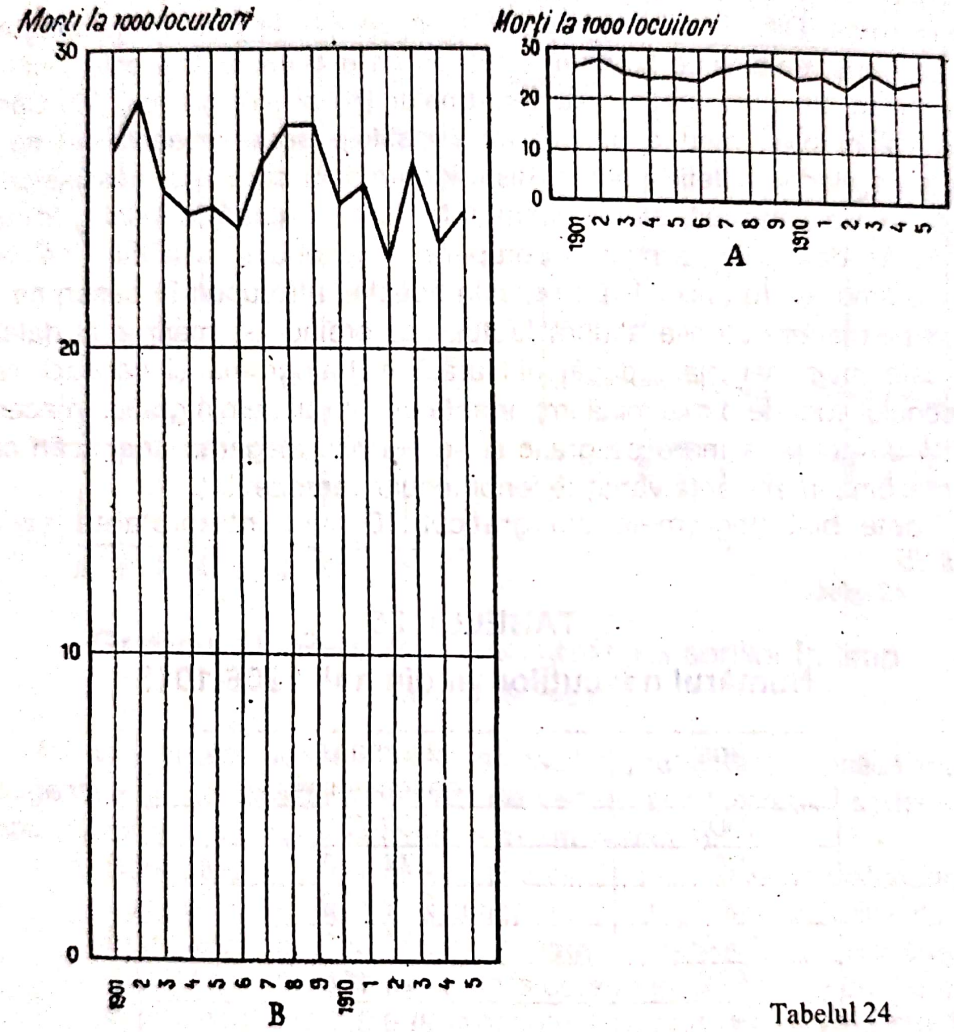
Extras din ***Anuarul statistic al României pe anul 1930****, p. 4.*

[1] *În cursul acestor paragrafe anticipăm puţin asupra unor lucruri ce vor fi expuse mai târziu. Credem totuşi că elementele la care facem anticipat apel sunt dintre acelea care fac parte din domeniul cunoştinţelor generale, aşa încât înlănţuirea expunerii nu va avea de suferit din această cauză.*

Scanned with OKEN Scanner

Această serie în timp poate da naștere fie la historiograma (A), fie la historiograma (B) (graficul 9).

Prima ne înfățișează fenomenul mortalității în România ca un fenomen de intensitate prea puțin variabilă de la un an la altul. Cu totul altă impresie ne face historiograma (B).



Tabelul 24

Graficul 9. Reprezentarea grafică pe scări diferite ale aceleiași serii în timp.

Posibilitatea aceasta, de a prezenta același fapt sub două aspecte direct opuse, constituie unul din mijloacele întrebuințate pentru falsificarea adevărului, pornind de la date statistice. Metoda statistică nu este însă de vină pentru aceasta. Institutul Internațional de Statistică a stabilit încă din anul 1870, cu prilejul celei de-a VIII-a sesiune a sa, în urma raportului dr.JACQUES BERTILLON, că unitățile trebuie astfel alese încât alura tangențială a liniei din

historiogramă să fie reprezentată de o dreaptă înclinată la 45° față de axe[1]. Vom vedea mai târziu care este mijlocul de a ajunge aici; deocamdată relevăm faptul că historiograma (B) este alcătuită după aceste principii[2].

Tot pentru a nu avea o impresie falsă despre mărimea variațiilor unui fenomen în raport cu însăși valoarea lui, este necesar ca întotdeauna axa absciselor să indice nivelul zero pentru valoarea ordonatelor și să fie figurată pe historiogramă. Diferența între impresiile pe care ni le face o historiogramă în care axa absciselor este figurată și alta în care această axă este lăsată de o parte reiese din comparația historiogramelor (A) și (B), graficul 10. Când, din cauza valorii prea mari a datelor ce constituie seria, aceasta nu se poate reprezenta clar și estetic printr-o historiogramă în care axa absciselor să fie figurată în mod obișnuit, atunci aceasta trebuie să apară în josul graficului, de care să fie despărțită printr-o întrerupere a desenului, așa cum se vede în historiograma (C) (graficul 10). Prezența acestei întreruperi în desen ne atrage în mod permanent atenția asupra faptului că ordinul de mărime al datelor din serie este mult mai mare decât ni-l arată historiograma și că deci variațiile fenomenului sunt de o mai mică importanță decât reiese din grafic. În acest mod se evită un gol prea mare pe grafic și se permite alegerea unei scări care să pună mai bine în evidență variațiile fenomenului reprezentat.

Toate historiogramele din graficul 10 sunt interpretarea seriei din tabelul 25.

TABELUL 25

Numărul născuților vii din anii 1906-1915

Anii	Numărul născuților vii
1906	262 438
1907	274 487
1908	272 850
1909	282 342
1910	273 106
1911	299 870
1912	314 090
1913	309 625
1914	327 345
1915	319 544

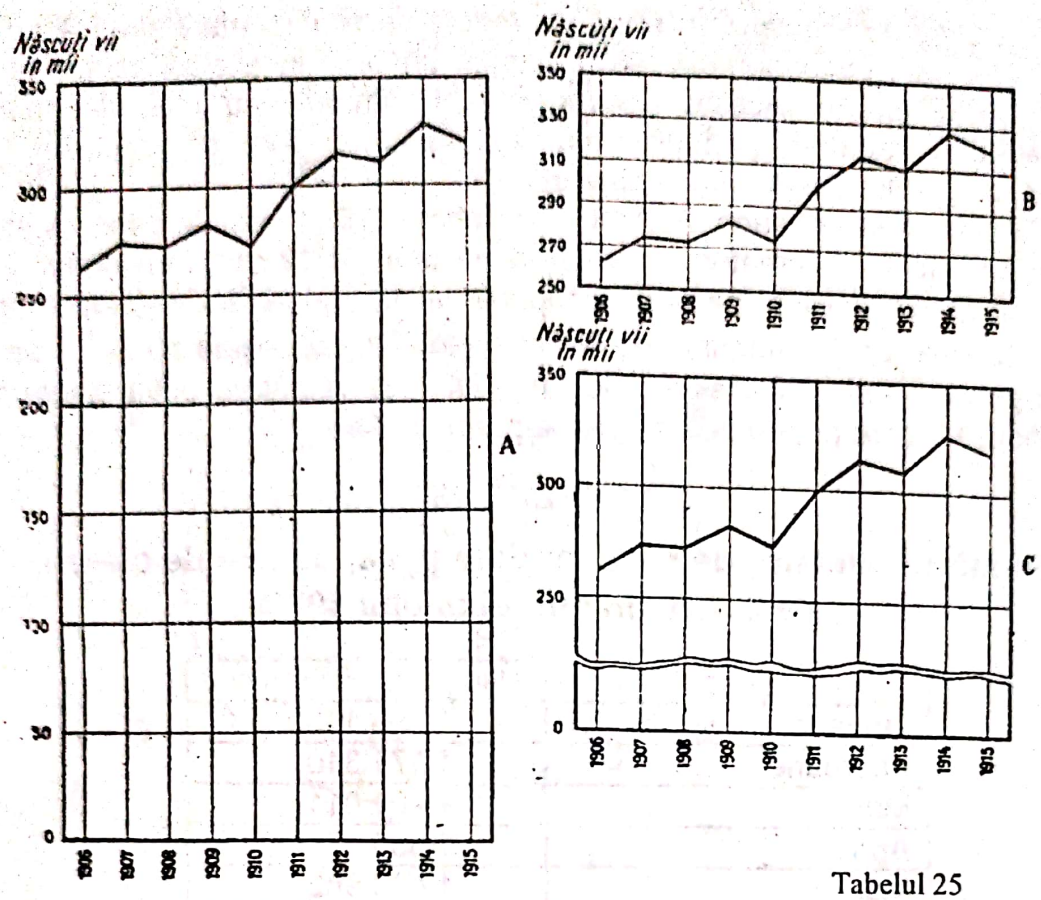
Extras din ***Anuarul statistic al României, 1930****, p. 41.*

---

[1] *Institut International de Statistique, VIII-e session, Dr.J. Bertillon,* ***Proposition relatives à l'uniformité dans l'établissement des graphiques****, p. 313.*

[2] *Vezi capitolul* Analiza seriilor în timp.

Scanned with OKEN Scanner



Tabelul 25

Graficul 10. Reprezentarea grafică a seriilor în timp și axa absciselor.

Mai sunt și alte norme menite să facă reprezentarea grafică mai clară, mai sugestivă și mai exactă din punct de vedere al interpretării seriei statistice corespunzătoare; cele mai importante sunt următoarele:

1. Sunt câteva categorii de linii care trebuie trase pe historiogramă, în așa fel încât să se distingă în mod precis de celelalte linii ce formează rețeaua de coordonate. Acestea sunt: linia care indică mersul fenomenului, linia care corespunde valorii 0 pentru ordonate și, în cazul unei serii de procente, linia care indică valoarea 100.
2. Nici o linie care nu e strict necesară nu trebuie trasă pe grafic. Se înșală cei ce cred că, prin mai multe detalii pe grafic, acesta își atinge mai ușor scopul său.
3. Cifrele relative la scara timpului și a ordonatelor trebuie scrise respectiv în josul și la stânga graficului. Când acesta este oarecum lung, atunci ultimele cifre trebuie scrise și la dreapta lui. Uneori poate fi util să se figureze pe grafic și valorile reprezentate, în care caz acestea se scriu pe marginea de sus.

Deși simple, aceste norme au importanța lor și dacă nu ne oprim mai mult asupra lor este din cauză că nu cer mai multă lămurire. A le avea în vedere la construcția unui grafic este lucrul principal.

Scanned with OKEN Scanner

29. *Diagrame polare.* În afară de modul de reprezentare grafică a unei serii în timp prin historiogramă, se utilizează uneori - în acelaşi scop - un tip anumit de diagrame. Aceste diagrame sunt întrebuinţate mai ales pentru reprezentarea variaţiilor periodice ale unui fenomen care este influenţat de anotimp, de poziţia zilei în săptămână etc.

Să considerăm seria în timp din tabelul 26. Această serie poate fi reprezentată în felul următor: se împarte un cerc în 12 sectoare egale şi se determină razele acestor sectoare proporţional cu datele seriei (diagrama A) sau pe 12 raze ce fac unghiuri egale între ele - şi deci egale cu 30° - se iau lungimi proporţionale cu aceste date (diagrama B). Astfel se obţin diagramele din graficul 11, care poartă numele de *diagrame polare.*

TABELUL 26

Numărul biletelor de 5 lei vândute în vagoanele de clasa I ale S.T.B., în cursul anului 1930

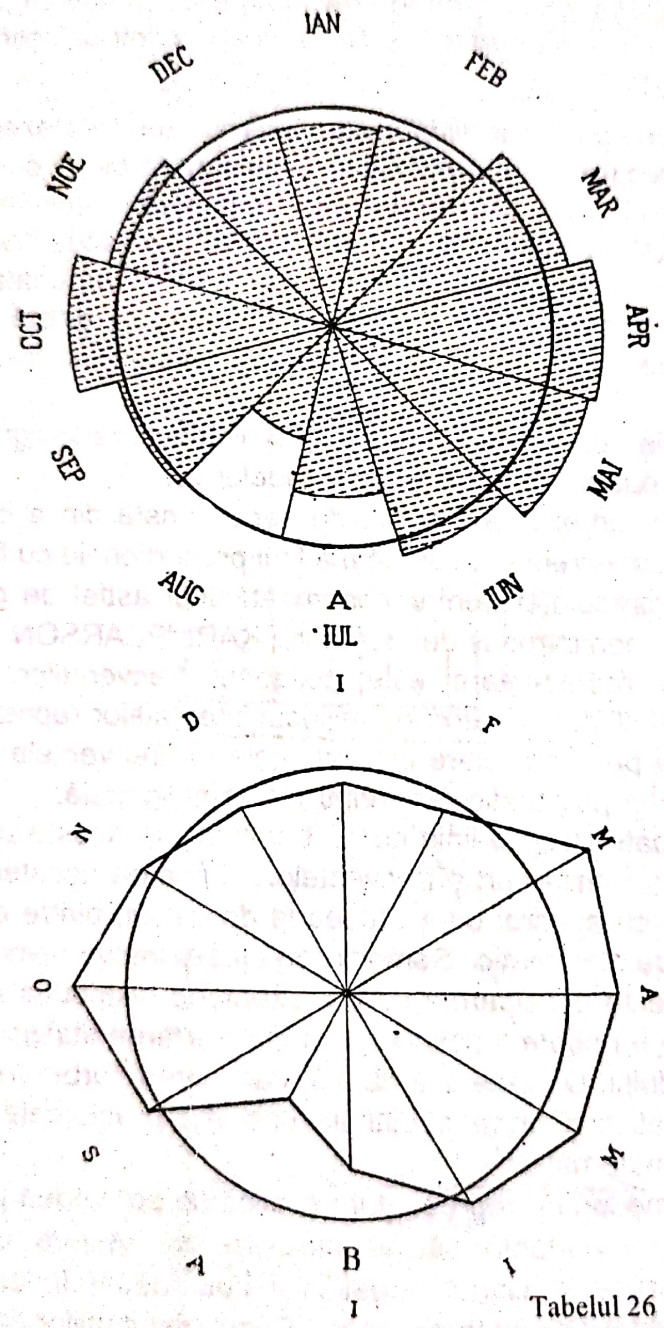
Luna	Numărul călătorilor
Ianuarie	1 150 115
Februarie	1 171 310
Martie	1 401 011
Aprilie	1 536 584
Mai	1 480 982
Iunie	1 345 855
Iulie	976 475
August	672 894
Septembrie	1 287 457
Octombrie	1 532 387
Noiembrie	1 312 922
Decembrie	1 156 562

Extras din ***Darea de seamă a Societăţii Comunale a Tramvaielor Bucureşti, pe anul 1930****, p.24.*

Diagramele polare au un singur avantaj: acela de a permite urmărirea în continuare a variaţiilor periodice ale unui fenomen, independent de momentul ales ca origine a timpului. În aceste diagrame, fiecare lună este precedată şi urmată de cele două linii ale anului în ordinea lor de succesiune cronologică. Totuşi ambele tipuri, A şi B, prezintă neajunsul de care am mai vorbit şi care constă din faptul că, elementele reprezentative nefiind la fel de înclinate faţă de o direcţie oarecare, ochiul este mai susceptibil de a comite erori în aprecierea raporturilor numerice[1]. În afară de aceasta, diagrama de tipul A prezintă un alt

[1] *Pe acest fapt se bazează construcţia mai tuturor jocurilor distractive cunoscute sub numele de «iluzii optice».*

mare dezavantaj: datele seriei sunt reprezentate pe o astfel de diagramă prin sectoare de cerc ale căror raze sunt proporționale cu ele; or, în acest caz, suprafețele care constituie în fond elementul reprezentativ sunt proporționale cu pătratul valorilor din serie, fapt care face ca valorile mai mari să fie exagerat reprezentate, în detrimentul celor mai mici.



Tabelul 26

Graficul 11. Numărul biletelor de 5 lei, vândute lunar în vagoanele S. T. B.

S-ar putea înlătura acest neajuns prin determinarea razelor sectoarelor în mod proporţional cu rădăcina pătrată a datelor, însă, în acest caz, diagrama ar pierde din suplețe şi poate tocmai din această cauză acest procedeu nu e consacrat.

Cu toate neajunsurile lor, diagramele polare sunt totuşi utilizate mai ales în lucrările de popularizare. Observăm în treacăt că este comod ca pe astfel de diagrame să reprezentăm valoarea medie a seriei printr-un cerc de rază proporţională cu aceasta.

30. *Reprezentarea grafică a distribuţiilor heterograde*. Reprezentarea grafică a distribuţiilor heterograde se bazează pe acelaşi principiu al coordonatelor carteziene care a servit şi la construcţia historiogramelor. Valorile variabilei independente, adică ale atributului în raport cu care s-a făcut distribuţia, se aşază pe axa absciselor; în dreptul lor se reprezintă în ordonată valoarea frecvenţelor corespunzătoare. O dată această operaţie făcută, există trei moduri diferite de a completa graficul, moduri care ne conduc la trei tipuri de reprezentări.

Să ne fixăm ideile deocamdată asupra unei distribuţii heterograde făcute pe o scară cu intervale egale, cum este cea din tabelul 12.

Primul mod de a reprezenta o astfel de serie constă din a construi pe fiecare din intervalele scării dreptunghiuri de înălţimi proporţionale cu frecvenţele distribuţiei. Termenul întrebuinţat pentru denumirea unui astfel de grafic este acela de *histogramă*, termen introdus de profesorul KARL PEARSON.

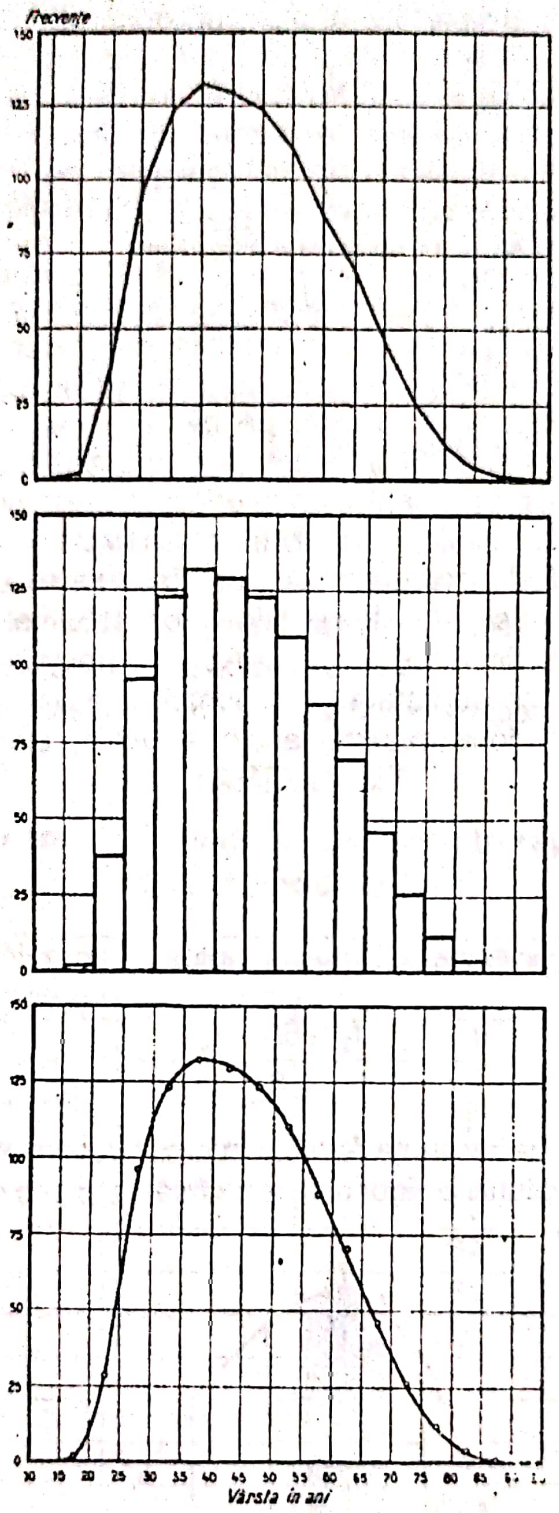
Un alt mod de reprezentare este poligonul frecvenţelor. Poligonul frecvenţelor se obţine în chipul următor: pe mijlocul intervalelor reprezentate pe axa absciselor se ridică perpendiculare proporţionale cu frecvenţele şi apoi se unesc extremităţile acestor perpendiculare printr-o linie poligonală.

În fine, se mai poate trage o linie curbă trecând prin aceste puncte sau cât mai aproape de ele, numită curba frecvenţelor. Trasarea acestei curbe se poate face din ochi sau cu ajutorul unor procedee de calcul, dintre care unele cer cunoştinţe speciale de matematici. Semnificaţia curbei frecvenţelor este de o importanţă capitală: ea este considerată ca reprezentând distribuţia ideală sau teoretică, distribuţie care nu poate fi obţinută prin înregistrarea statistică directă, din cauza jocului hazardului. Din această cauză, calcularea curbei frecvenţelor pornind de la datele statistice brute constituie una din problemele cele mai importante ale statisticii matematice[1].

Toate aceste trei moduri de reprezentare grafică se pot vedea pe graficul 12. Este de neapărată importanţă să nu pierdem din vedere că aceste reprezentări nu se pot face în chipul arătat mai sus, decât în cazul când *distribuţia este făcută în raport cu intervale egale*. Cazul distribuţiilor cu intervale neegale cere o tratare specială şi diferită de precedenta.

[1] *Vezi* Cap. VI.

Scanned with OKEN Scanner



A = poligonul frecvențelor.
B = histogramă.
C = curba frecvențelor.

Tabelul 12

Graficul 12. Cele trei moduri de reprezentare a unei distribuții heterograde.

Scanned with OKEN Scanner

Dintre aceste trei tipuri de grafice, poligonul și curba frecvențelor sunt întrebuințate mai ales în scopul de a degaja *forma* unei distribuții sau pentru a face o comparație grafică a două serii. Histograma pune în evidență mai bine decât tipurile precedente diferența de mărime între frecvențe, nu degajează însă așa de bine forma distribuției; este întrebuințată mai ales în cazurile când distribuția conține intervale cu frecvențe de mărime relativ mică alături de altele cu frecvențe de un ordin de mărime mult superior.

31. *Frecvențele și suprafața poligonului frecvențelor.* Suprafața închisă de linia în scară ce mărginește histograma și axa absciselor este proporțională cu numărul total al frecvențelor. Același lucru se poate spune și de poligonul frecvențelor, cu o singură deosebire: pe când în cazul histogramei această proporționalitate între suprafețe și frecvențe se menține pentru o parte oarecare a distribuției, eventual numai pentru un interval, în cazul poligonului frecvențelor proporționalitatea nu este valabilă decât în ansamblu.

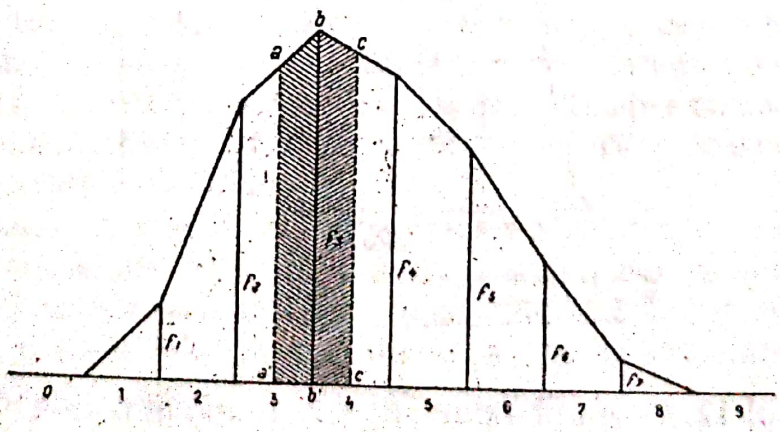
Într-adevăr, să descompunem suprafața cuprinsă între un poligon de frecvență și axa absciselor în elementele ei constitutive: două triunghiuri și o serie de trapeze (graficul 13). Fie f_1, f_2,..., f_n frecvențele corespunzătoare poligonului. Se vede că suprafața corespunzătoare unui interval, de exemplu 3-4, este formată din două trapeze, *aba'b'* și *bcb'c'*, ale căror suprafețe sunt respectiv egale cu:

$$\frac{f_2 + 3f_3}{8} \qquad \frac{3f_3 + f_4}{8}.$$

Deci, suprafața corespunzătoare intervalului 3-4 fiind:

$$\frac{f_2 + 6f_3 + f_4}{8}$$

nu este proporțională cu frecvența f_3 decât în cazul când $f_2 + f_4 = 2f_3$. Acest fapt se întâmplă numai când laturile poligonului ce se întâlnesc în *b* sunt în prelungire.



Graficul 13.

Scanned with OKEN Scanner

Însumând acum toate suprafețele, obținem suprafața totală:

$$\frac{f_1}{2}+\frac{f_1+f_2}{2}+\frac{f_2+f_3}{2}+\ldots+\frac{f_6+f_7}{2}+\frac{f_7}{2}=f_1+f_2+f_3+\ldots+f_7$$

care reprezintă într-adevăr totalul frecvențelor.

Din cauza faptului că și histograma, și poligonul frecvențelor închid suprafețe proporționale cu totalul frecvențelor, rezultă un principiu care trebuie să ne călăuzească în trasarea din ochi a curbei frecvențelor. Trasând această curbă de-a lungul poligonului sau histogramei, observăm că uneori, în mersul ei, curba adaugă suprafețe, alteori diminuează suprafața închisă de primele; e bine ca trasarea să fie făcută în așa fel încât totalul suprafețelor adăugate să fie egal cu totalul celor ce se scad.

32. Când distribuția este făcută în raport cu o variabilă discontinuă, cele trei tipuri de grafice examinate, deși servesc destul de bine și în acest caz scopului urmărit prin reprezentarea grafică, trebuie pe cât posibil evitate. Reprezentarea grafică a unei astfel de distribuții printr-un poligon sau curbă de frecvențe ne-ar da impresia continuității variabilei independente.

TABELUL 27

Distribuția florilor de *Pimpinella saxifraga* în raport cu numărul razelor pedunculului

Numărul razelor pedunculului	Numărul florilor
5	1
6	5
7	9
8	22
9	38
10	62
11	61
12	29
13	14
14	4
15	4
	249

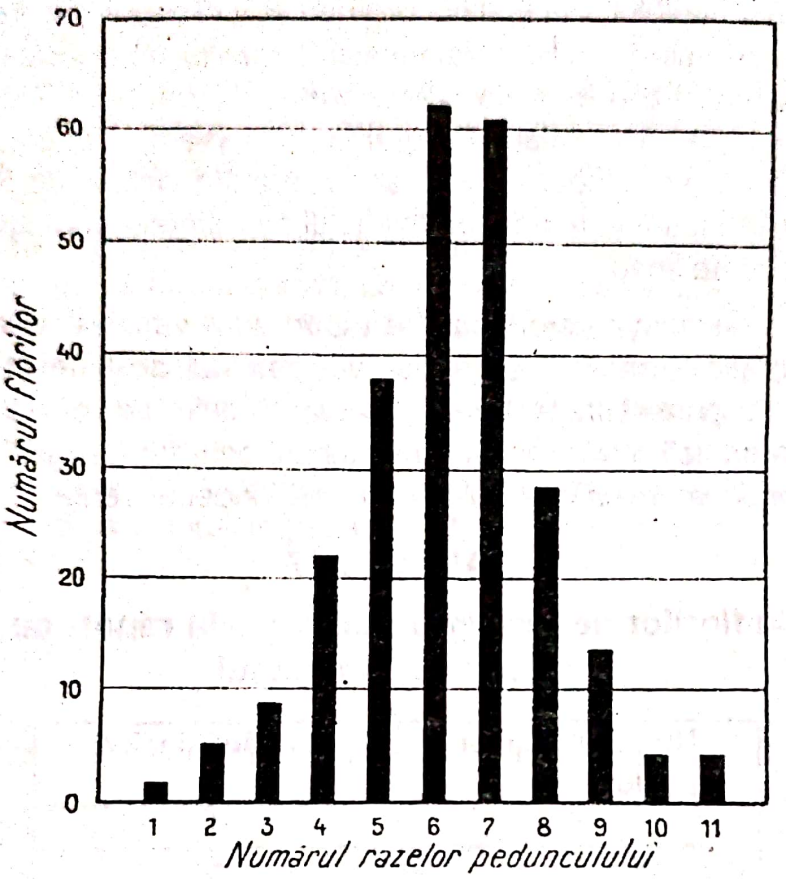
Reprodus după A. Julin, ***op. cit.****, p. 358.*

Pentru astfel de distribuții, cum este cea din tabelul 27, e preferabil să întrebuințăm reprezentarea prin bare sau, în orice caz, să recurgem la

Scanned with OKEN Scanner

histogramă. Reprezentarea grafică corespunzătoare acestui tabel este cea înfățișată de graficul 14.

Nu trebuie scăpat din vedere că, în cazul reprezentării grafice a unei distribuții discontinue printr-o histogramă, dreptunghiurile ce o alcătuiesc trebuie astfel construite încât mijlocul bazei lor să coincidă cu valorile scării distribuției.



Tabelul 27

Graficul 14. Reprezentarea grafică a distribuțiilor heterograde de variabilă discontinuă.

33. *Intervale neegale*. Procedeele de reprezentare grafică expuse până acum se aplică numai distribuțiilor cu intervale egale, fapt care nu trebuie pierdut niciodată din vedere.

Într-adevăr, distribuțiile heterograde ale căror intervale nu sunt de aceeași mărime cer o transformare specială înainte de a putea fi reprezentate grafic într-un mod judicios.

Să considerăm o astfel de distribuție cum este cea din tabelul 28.

Observăm că primele trei intervale sunt egale cu 100 locuitori, al patrulea cu 200 locuitori și, în fine, ultimele cu 500. Presupunând că avem de reprezentat o astfel de distribuție printr-o histogramă, vom ține seamă de faptul că pe un atare grafic frecvențele distribuției vor fi reprezentate prin suprafețe propor-

Scanned with OKEN Scanner

ţionale cu ele. Prin urmare, este necesar ca dreptunghiul reprezentând frecvenţa corespunzătoare intervalului 300-499, şi care avea deci o bază de două ori mai mare decât dreptunghiurile intervalelor precedente, să aibă o înălţime proporţională nu cu frecvenţa 129, ci cu jumătatea acesteia, 64,5. La fel frecvenţele intervalelor egale cu 500 de locuitori vor fi reprezentate prin dreptunghiuri de înălţimi proporţionale cu a cincea parte a lor.

TABELUL 28

Distribuţia satelor din judeţul Gorj, în raport cu numărul locuitorilor

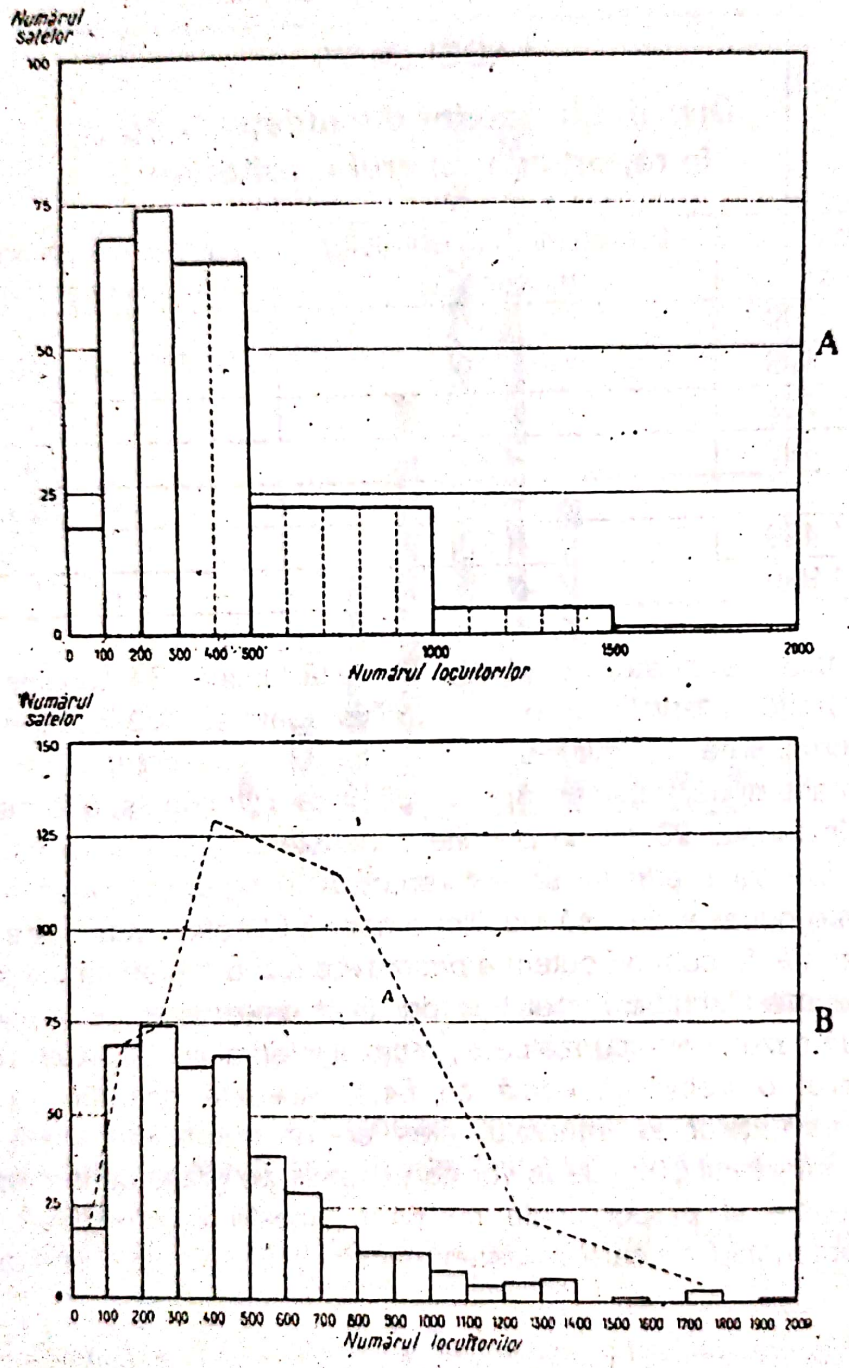
Numărul locuitorilor	Frecvenţele intervalelor menţionate	Frecvenţele reduse la intervalul de 100 locuitori
0 - 99	19	19
100 - 199	69	69
200 - 299	74	74
300 - 499	129	64,5
500 - 999	114	22,8
1 000 - 1 499	23	4,6
1 500 - 1 999	5	1

Acest fapt mai poate fi privit şi sub o altă lumină. Să considerăm iarăşi intervalul 300-499. Intervalele precedente fiind egale cu 100 locuitori, e nevoie pentru uniformizarea intervalelor distribuţiei să descompunem intervalul considerat în alte două, egale fiecare cu 100. Însă distribuţia sub forma sub care se prezintă în tabelul 28 nu ne permite cunoaşterea modului în care cazurile încadrate în intervalul primitiv se vor repartiza în raport cu noua împărţire a acestuia[1]. Este necesar deci să stabilim o normă teoretică după care să facem această repartiţie. Şi cum nu putem *a priori* avea nici o preferinţă pentru vreunul din noile intervale, lucrul cel mai bun de făcut este să atribuim acestora un număr egal de cazuri corespunzătoare. În cazul intervalului 300-499, vom atribui fiecărui interval o frecvenţă egală cu 64,5. Această operaţie se numeşte reducerea frecvenţelor la intervalul ales ca unitate. Celor două intervale provenite din intervalul 300-499 le vor corespunde pe grafic două dreptunghiuri de înălţimi egale şi proporţionale cu 64,5. Aceste dreptunghiuri contopite formează dreptunghiul de care vorbeam mai sus şi care corespunde intervalului

[1] *Cum tabelul 28 este calculat cu ajutorul datelor tabelului 15 B, se întâmplă în cazul de faţă că această repartiţie este cunoscută. Nu acesta este cazul care se întâlneşte în practică. Dacă am folosit pentru exemplificarea procedeelor de reprezentare grafică a distribuţiilor heterograde cu intervale neegale o distribuţie dedusă dintr-alta cu intervalele egale, am făcut-o în interesul priceperii mai lesnicioase a diferenţei ce există între aceste procedee de la un caz la altul.*

Scanned with OKEN Scanner

300-499. Așa că, oricum ar fi privită chestiunea reprezentării grafice printr-o histogramă a unei distribuții cu intervale egale, graficul care se obține este același și nici nu putea fi altfel, deoarece realitatea nu poate avea decât o singură formă.



Tabelul 15 B și 28

Graficul 15. Reprezentarea grafică a unei distribuții heterograde în raport cu intervale egale și neegale.

Poligonul frecvențelor, deși acest mod de reprezentare a distribuțiilor nu ar fi recomandat în cazul când intervalele lor nu sunt egale, se obține unind mijloacele bazelor superioare ale dreptunghiurilor reprezentând frecvențele reduse. Acest fapt face ca pe alocuri linia poligonală să se confunde cu histograma. Orice alt mod de a trasa poligonul frecvențelor distruge proporționalitatea între totalul frecvențelor și suprafața închisă de el.

Reprezentarea grafică a distribuției din tabelul 28 este înfățișată în graficul 15 A. Dăm cu titlu de exemplu și reprezentarea grafică a aceleiași distribuții sub forma pe care ar avea-o dacă nu s-ar face în prealabil reducerea frecvențelor la intervale egale. Și în același timp dăm și reprezentarea grafică a aceluiași fenomen în forma sub care se prezintă distribuția din tabelul 15 B. Aceasta, pentru a se face o comparație între felurile cum ne sunt relevate caracterele unei distribuții în cele două cazuri distincte: în raport cu o scară cu intervale egale și în raport cu una cu intervale neegale.

34. *Forma distribuțiilor heterograde.* Caracterul cel mai important pus în evidență prin reprezentarea grafică a unei distribuții este forma acesteia. Studiul formei unei distribuții constituie o chestiune de prim ordin în statistica matematică. De aceea s-a căutat ca acest lucru să se facă nu numai prin intermediul reprezentării grafice - procedeu oarecum subiectiv -, ci și prin mijloace de calcul care au avantajul de a înlătura orice apreciere personală. Profesorul Karl Pearson a elaborat în această direcție o întreagă disciplină care însă e departe de a putea fi expusă complet într-o lucrare de felul celei de față.

Vom rămâne la studiul grafic al formei distribuțiilor și din această cauză nu vom merge prea departe cu diferențierea tipurilor ce s-ar putea prezenta, oprindu-ne la următoarele cinci:

1. Distribuție simetrică (profil de clopot);
2. Distribuție ușor asimetrică,
3. Distribuție pronunțat asimetrică;
4. Distribuție în formă de J (*J-shaped*);
5. Distribuție în formă de U (*U-shaped*).

Această clasificare a distribuțiilor în raport cu forma lor s-a făcut ținând seamă de poziția pe care o ocupă în scara distribuției intervalul căruia îi corespunde frecvența maximă.

35. *Distribuții simetrice.* În prima categorie intră distribuțiile care, reprezentate prin poligonul de frecvență corespunzător, se prezintă sub forma profilului unui clopot. Intervalul de frecvență maximă se găsește la mijlocul scării distribuției, diminuarea frecvențelor spre dreapta și stânga acestui interval făcându-se în același ritm. Desigur că se mai pot imagina și alte forme de curbe simetrice, totuși, cu excepția unora ce fac parte din categoria a cincea, tipul descris mai sus este singurul care se prezintă mai des și în mod natural în studiul fenomenelor colective. Deși - ținând seamă de mulțimea de cazuri când distribuțiile se prezintă sub una din celelalte forme - putem spune că distribuția simetrică este un caz comparativ rar în statistică.

Scanned with OKEN Scanner

Din această categorie face parte celebra distribuție normală a lui LAPLACE-GAUSS. Distribuții normale se găsesc mai ales în biometrie și printre măsurile antropometrice. Tot normal se distribuie și diferitele date obținute cu ocazia unei măsurători, observații, percum și abaterile rezultate din tragerea la țintă.

Cazul tipic de distribuție normală este alcătuit de distribuția înălțimilor dintr-o populație omogenă.

Tabelul 29 ne înfățișează o astfel de distribuție, iar graficul 16 constituie reprezentarea grafică corespunzătoare.

TABELUL 29

Distribuție de talii a unui grup de adulți născuți în Regatul Unit

Înălțimea în degete[1]	Anglia	Scoția	Țara Galilor	Irlanda	Total
57-	1	-	1	-	2
58-	3	1	-	-	4
59-	12	-	1	1	14
60-	39	2	-	-	41
61-	70	2	9	2	83
62-	128	9	30	2	169
63-	320	19	48	7	394
64-	524	47	83	15	669
65-	740	109	108	33	990
66-	881	139	145	58	1 223
67-	918	210	128	73	1 329
68-	886	210	72	62	1 230
69-	753	218	52	40	1 063
70-	473	115	33	25	646
71-	254	102	21	15	392
72-	117	69	6	10	202
73-	48	26	2	3	79
74-	16	15	1	-	32
75-	9	6	1	-	16
76-	1	4	-	-	5
77-	1	1	-	-	2
Total	**6 194**	**1 304**	**741**	**346**	**8 585**

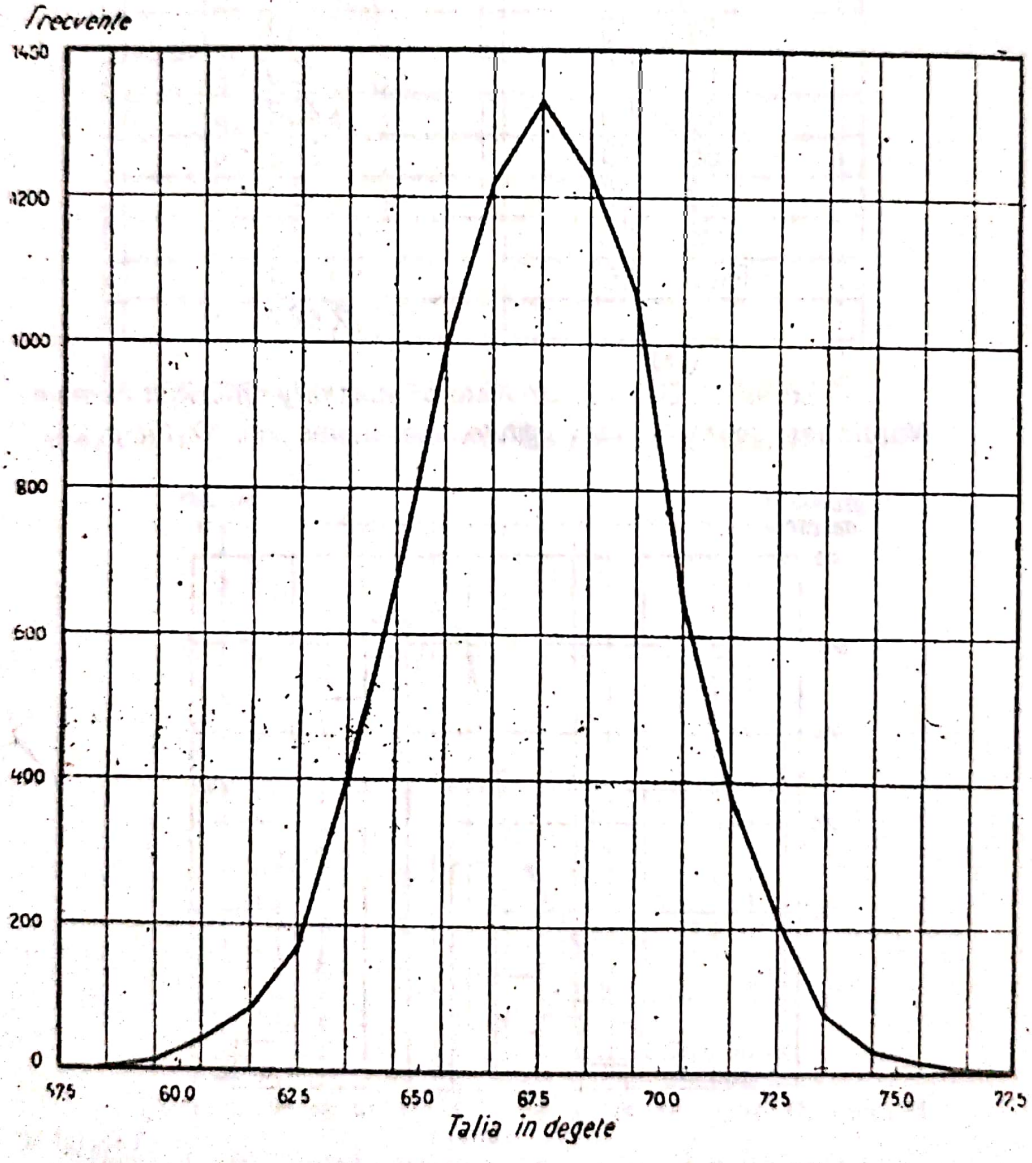
*Reprodus după G. U. Yule, **op. cit.**, p. 88.*

[1] *Un deget (inch)= 2,56 cm.*

Scanned with OKEN Scanner

36. *Distribuții ușor asimetrice*. Dacă descreșterea frecvențelor nu se face la fel de repede de o parte și de alta a intervalului de frecvență maximă, fără ca diferența între aceste moduri de descreștere a frecvențelor să fie prea mare, distribuția se numește ușor asimetrică. Tipul acesta de distribuție este cel care se întâlnește cel mai des în statistică și putem spune că nu există domeniu în care această metodă își găsește aplicații și în care să nu întâlnim tipuri de distribuții ușor asimetrice.

Din această categorie fac parte mai ales distribuțiile fenomenelor în legătură cu evoluția vieții. Distribuția bărbaților căsătoriți în raport cu vârsta, întâlnită mai sus, este un exemplu de distribuție ușor asimetrică (tabelul 12, graficul 12).



Tabelul 29

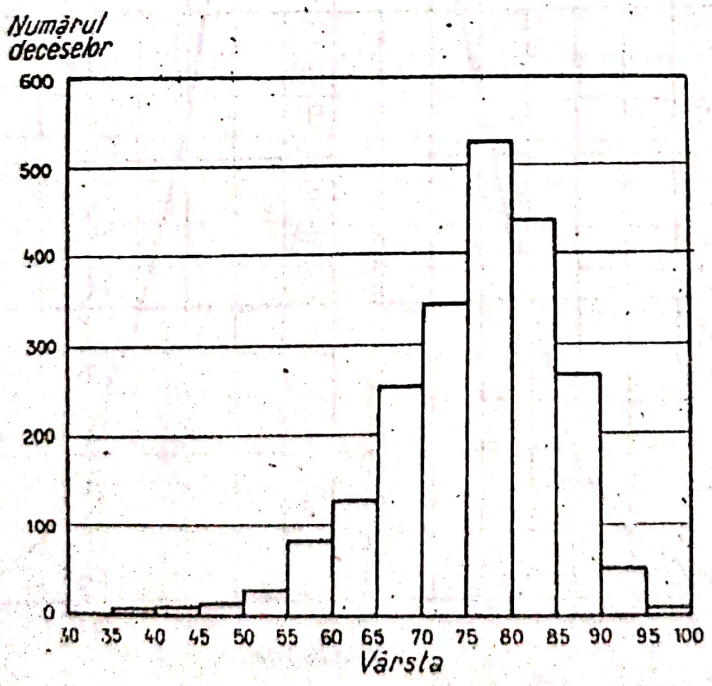
Graficul 16. Distribuție de talii.

TABELUL 30

Mortalitatea unui grup de funcționare

Vârsta	Decese
30 - 35	1
35 - 40	5
40 - 45	8
45 - 50	12
50 - 55	28
55 - 60	82
60 - 65	128
65 - 70	253
70 - 75	342
75 - 80	525
80 - 85	438
85 - 90	265
90 - 95	53
95 - 100	18
Peste 100	4
	2 162

Extras din G. King, ***On Rate of Mortality amongst Female Nominees****, Journal of the Institute of Actuaries, vol. XXXIII, p. 262.*



Tabelul 30

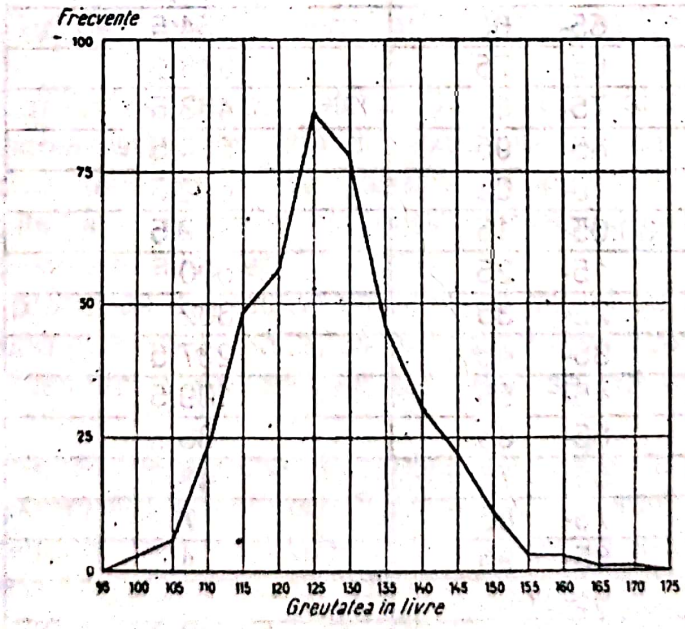
Graficul 17. Distribuția deceselor unui grup de funcționare, în raport cu vârsta.

TABELUL 31

Distribuția unui grup de bărbați între 20-24 ani, de înălțime egală cu 5 picioare și 3 degete, în raport cu greutatea corpului

Greutatea în livre[1]	Frecvențe
97,5 - 102,5	3
102,5 - 107,5	6
107,5 - 112,5	23
112,5 - 117,5	48
117,5 - 122,5	56
122,5 - 127,5	86
127,5 - 132,5	78
132,5 - 137,5	45
137,5 - 142,5	30
142,5 - 147,5	22
147,5 - 152,5	11
152,5 - 157,5	3
157,5 - 162,5	3
162,5 - 167,5	1
167,5 - 172,5	1
	416

Extras din H. L. Rietz, ***Handbook of Mathematical Statistics****, p.21.*



Tabelul 31

Graficul 18. Distribuție heterogradă ușor asimetrică.

[1] *O livră sau un fund = 453 g.*

Același tip de distribuție este cea din tabelul 30 (graficul 17).

Sunt cazuri când această distribuție este întâlnită și printre elementele biometrice. Greutatea corpului se distribuie de obicei după o curbă ușor asimetrică (tabelul 31, graficul 18).

În fine, pentru a ilustra și mai bine varietatea fenomenelor ce se distribuie ușor asimetric, dăm și distribuția înălțimii barometrice după datele din localitatea Southampton (tabelul 32, graficul 19).

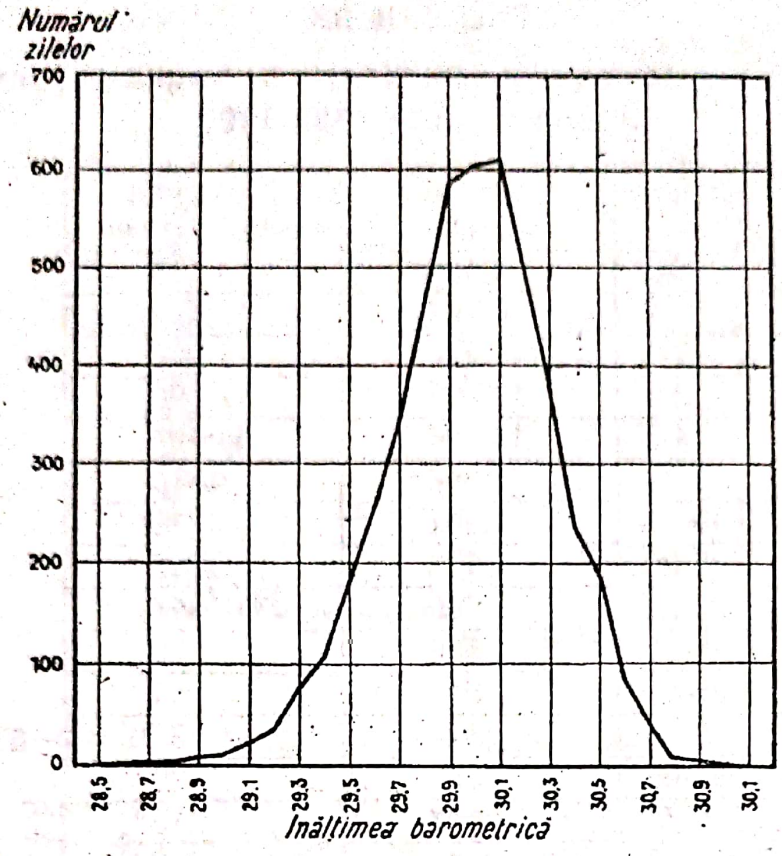
TABELUL 32

Distribuția înălțimii barometrice la Southampton, în cursul anilor 1878-1890

Înălțimea în degete	Numărul zilelor
28.45-28.55	1
55- 65	2
65- 75	2
75- 85	4
85- 95	8,5
95-29.05	13,5
29.05- 15	21,5
15- 25	37
25- 35	79
35- 45	108
45- 55	181,5
55- 65	254,5
65- 75	348,5
75- 85	463,5
85- 95	584,5
95-30.05	602,5
30.05- 15	619,5
15- 25	500
25- 35	382
35- 45	237,5
45- 55	189,5
55- 65	88,5
65- 75	43,5
75- 85	7
85- 95	4
95-31.05	1
	4 784

Extras din K. Pearson and A. Lee, ***Philosophical Transactions****, A, vol. CXC, p. 428.*

Scanned with OKEN Scanner



Tabelul 32

Graficul 19. Distribuția înălțimii barometrice la Southampton.

37. *Distribuții pronunțat asimetrice*. Dacă cea mai mare parte a frecvențelor este concentrată către una din extremitățile scării, în așa fel încât intervalul de frecvență maximă se găsește cu totul deplasat spre această extremitate, distribuția se numește pronunțat asimetrică. Distribuția din tabelul 33 (graficul 20) ilustrează acest tip.

Distribuții pronunțat asimetrice se prezintă destul de rar în practică. Se mai întâmplă, după cum vom vedea, că din cauza unei scări cu intervale prea mari, multe din distribuțiile din această categorie se pot prezenta sub forma distribuțiilor în J.

Ceea ce trebuie remarcat la distribuțiile asimetrice în general este faptul că cele două extremități ale curbei frecvențelor nu sunt egale. Traducând termenul englezesc, le vom numi *cozi*; vom distinge, așadar, o coadă lungă și una scurtă. Coada lungă poate să fie, după caz, de partea valorilor superioare ale scării distribuției sau de partea celor inferioare.

Poziția acesteia este de o importanță capitală în studiul fenomenelor colective. Coada lungă reprezintă întrucâtva elementele excepționale ce se întâlnesc în masa unităților statistice ce formează obiectul cercetărilor respective.

Scanned with OKEN Scanner

TABELUL 33

Distribuţia deceselor cauzate de difterie în Anglia şi Ţara Galilor, în cursul anilor 1891-1900

Vârsta	Numărul morţilor	Frecvenţe reduse la 1 an
0-1	4.186	4.186
1-2	10.491	10.491
2-3	11.218	11.218
3-4	12.390	12.390
4-5	11.194	11.194
5-10	23.348	4.670
10-15	4.092	818
15-20	1.123	225
20-25	585	117
25-35	786	79
35-45	512	51
45-55	324	32
55-65	260	26
65-75	127	13
Peste 75	35	?
	80 671	

Reprodus după G. U. Yule, ***op. cit.****, p. 98.*

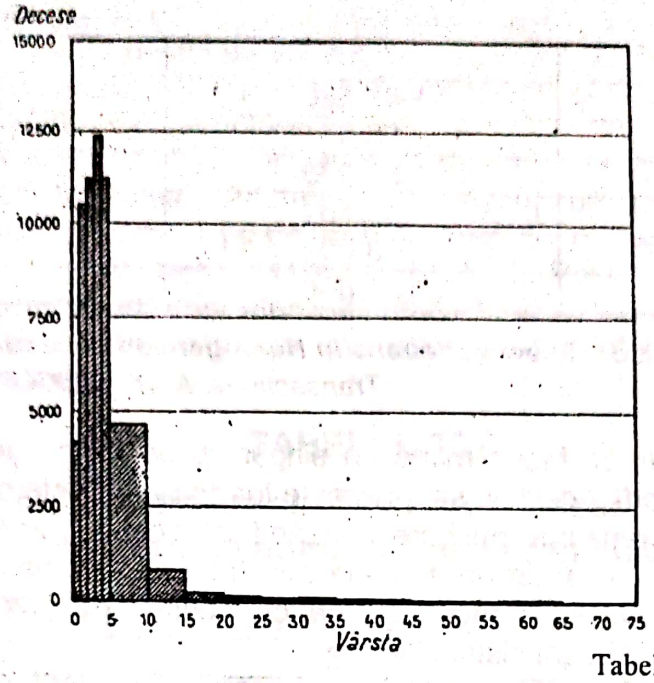
38. *Distribuţii în J.* Distribuţiile pentru care intervalul de frecvenţă maximă este chiar unul din intervalele extreme, frecvenţele celorlalte intervale mergând în descreştere către cealaltă extremitate a scării distribuţiei, poartă numele, după forma pe care o are curba frecvenţelor corespunzătoare, de distribuţii în formă de J.

Distribuţiile de venituri, de averi sunt exemple tipice ale acestei categorii. Importanţa acestui tip a fost în special pusă în evidenţă de economistul VILFREDO PARETO; de aceea, curba respectivă este cunoscută sub numele de *curba lui Pareto.*

Dăm mai jos datele lui Goschen asupra clădirilor din Londra, clasificate după valoarea lor fiscală.

Foarte multe din distribuţiile în formă de J ascund o distribuţie pronunţat asimetrică. Astfel, chiar distribuţia din tabelul 34 (graficul 21) poate să apară, după o analiză aprofundată, ca o distribuţie foarte asimetrică; e foarte puţin probabil ca frecvenţa maximă să corespundă intervalului 0-1 în cazul unei descompuneri a intervalului 0-10 în altele zece. Distribuţia din tabelul 33 ar fi apărut ca o distribuţie în formă de J, dacă primele 5 intervale ar fi fost contopite în unul singur.

Scanned with OKEN Scanner



Tabelul 33

Graficul 20. Mortalitatea de difterie în Anglia şi Ţara Galilor.

Cunoaşterea faptului dacă o distribuţie este sau nu în formă de J este o chestiune din cele mai delicate, deoarece - mai ales în cazul unei distribuţii continue - divizarea intervalului marginal, cel care interesează în acest caz, poate merge foarte departe. Şi cel puţin în mod teoretic, problema prezintă o latură faţă de care prudenţa cercetătorului nu poate fi niciodată prea mare. Dificultăţile dispar complet în cazul unei distribuţii discontinue, unde divizarea intervalelor este limitată în mod natural. Atunci când fiecare clasă a distribuţiei este reprezentată printr-o singură valoare a variabilei independente, putem spune cu certitudine dacă distribuţia are sau nu forma de J. Acesta este cazul distribuţiei în tabelul 35 (graficul 22).

TABELUL 34

Distribuţia caselor din Londra în raport cu valoarea lor fiscală

Valoarea fiscală în £	Numărul caselor	Frecvenţe reduse la intervalul de 10 £
1	2	3
0-10	3 174 806	3 174 806
10-20	1 450 781	1 450 781
20-30	441 595	441 595
30-40	259 756	259 756
40-50	150 968	150 968
50-60	90 432	90 432
60-80	104 128	52 064

Scanned with OKEN Scanner

(continuare tab. 34)

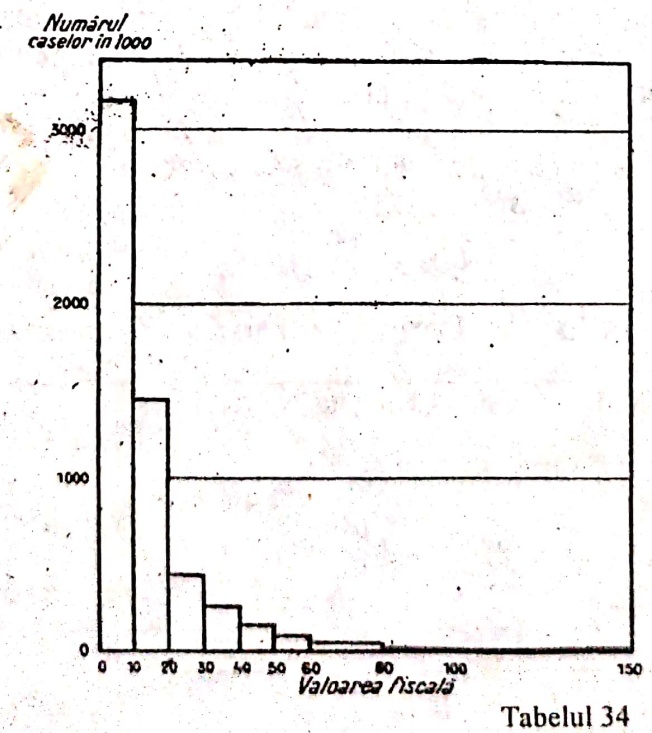
1	2	3
80-100	47 326	23 663
100-150	58 871	11 774
150-300	37 988	2 532
300-500	8 781	420
500-1 000	3 002	60
1 000-1 500	1 036	21

Extras din K. Pearson, ***Skew Variations în Homogeneous Material,*** *Philosophical Transactions, A. vol. CLXXXVI, 1895, p. 343.*

39. *Distribuţii în U.* Mai rămâne un singur tip de distribuţie: distribuţia în formă de U *(U- shaped)*. Acest tip se întâlneşte foarte rar. El este caracterizat prin prezenţa a două frecvenţe maxime corespunzând celor două intervale marginale.

Distribuţiile din acest tip pot fi şi ele simetrice sau asimetrice, după cum descreşterea frecvenţelor se face sau nu cu acelaşi pas de la intervalele marginale către centrul scării distribuţiei.

Tabelul 36 (graficul 23) constituie un exemplu de distribuţie din această categorie. Este vorba despre gradul de acoperire a cerului. Acest fenomen este aproape singurul care se distribuie în mod indiscutabil în formă de U. Faptul a fost pus pentru prima oară în evidenţă de prof. Karl Pearson în articolul *Cloudiness: Note on a Novel Case of Frequency* (***Proceedings of the Royal Society***, A. vol.CXCVII, 1901, p.443), de unde am extras şi datele tabelului 36.



Tabelul 34

Graficul 21. Distribuţia caselor din Londra în raport cu valoarea fiscală.

40. În afară de tipurile de distribuţii descrise în paragrafele precedente, mai există şi alte tipuri - destul de variate ca formă, dar nu tot aşa de frecvente - dintre care nu vom releva decât pe acelea care rezultă din suprapunerea a două populaţii complet diferite ca structură.

Distribuţia din tabelul 37 constituie un exemplu de astfel de suprapuneri de colectivităţi diferite. Graficul 24 ne pune în evidenţă prezenţa a două clase de frecvenţe maxime, fapt care poate fi interpretat ca o suprapunere de două distribuţii de tipul celor deja examinate. Prof. K. Pearson a pus problema posibilităţii descompunerii unei atare distribuţii în cele două forme suprapuse. Soluţia acestei probleme constituie una din cele mai interesante aplicaţii ale statisticii matematice[1].

TABELUL 35

Distribuţia femeilor din Fall River (Massachusetts) în raport cu numărul copiilor în viaţă

Numărul copiilor	Numărul femeilor	Frecvenţe relative
0	3 381	145
1	3 277	141
2	3 931	126
3	2 538	109
4	2 067	89
5	1 761	76
6	1 564	67
7	1 256	54
8	1 125	48
9	933	40
10	777	33
11	573	25
12	413	18
13	279	12
14	178	8
15	105	4
16	47	2
17	38	2
18	22	1
19	8	0
20	3	0
21	3	0
	23 279	**1 000**

Extras din E. E. Day, ***Statistical Analysis****, New-York, Macmillan, 1930, p.128.*

[1] *K. Pearson,* ***On the Dissection of Asymetrical Frequency Curves****, Philosophical Transactions, vol. 185, serie A, p.13.*

Scanned with OKEN Scanner

TABELUL 36

Distribuția gradului de acoperire a cerului la Breslau, în cursul anilor 1876-1885

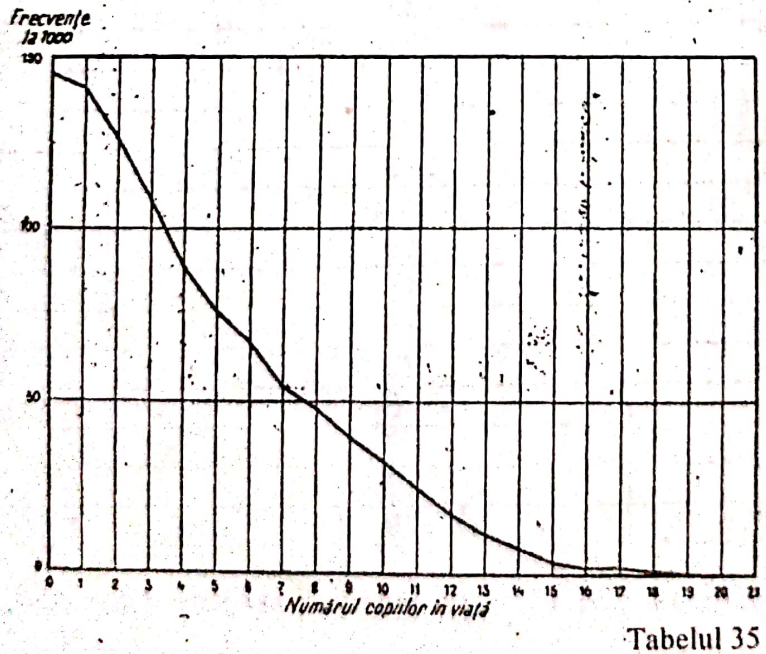
Gradul de acoperire	Numărul zilelor
0	751
1	179
2	107
3	69
4	46
5	9
6	21
7	71
8	194
9	117
10	2.089
	3.653

TABELUL 37

Distribuția unor spice de orz în raport cu procentul grăunțelor sterile

Procente	Frecvențe
0-5	76
5-10	131
10-15	105
15-20	87
20-25	14
25-30	53
30-35	128
35-40	147
40-45	135
45-50	70
50-55	28
55-60	11
60-65	6
65-70	6

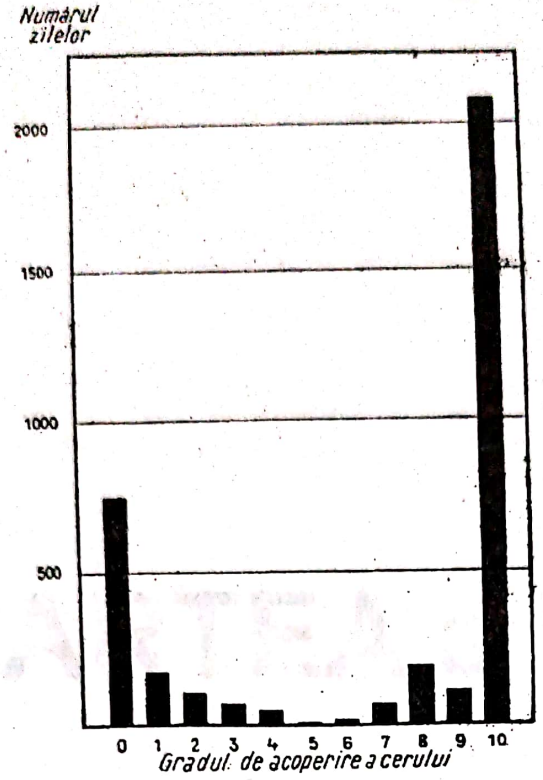
Reprodus din R. Risser și C. Traynard, ***Les principes de la statistique mathématique****, p. 100.*



Tabelul 35

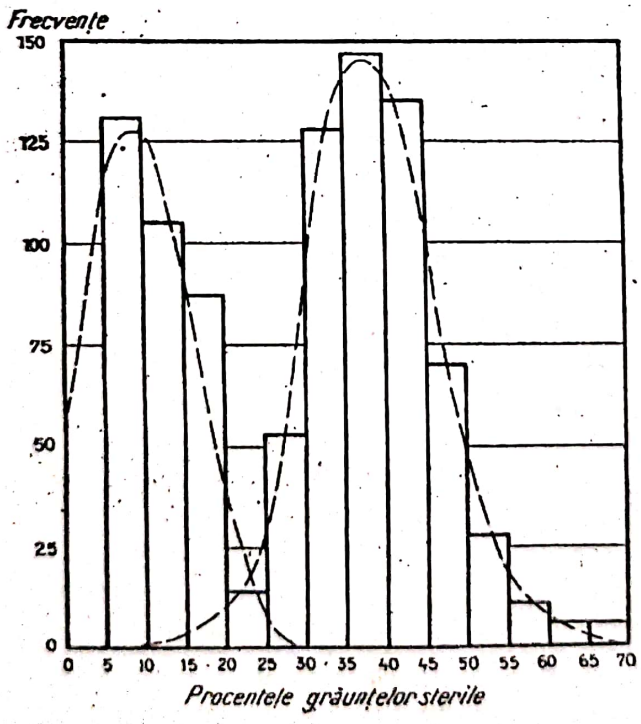
Graficul 22. Distribuția femeilor din Fall River, după numărul copiilor în viață.

Scanned with OKEN Scanner



Tabelul 36

Graficul 23. Gradul de acoperire a cerului, la Breslau.



Tabelul 37

Graficul 24. Distribuție de colectivități suprapuse.

Scanned with OKEN Scanner

PARTEA a II-a

Scanned with OKEN Scanner

Capitolul IV

Valori tipice

41. *Generalități.* Clasificarea unei serii de valori observate, făcută sub forma de distribuție heterogradă, a avut ca scop, după cum am accentuat la timp, prezentarea varietății primitive sub o formă mai ușor de considerat. Spiritul omenesc nu poate să prindă dintr-o dată ceea ce este esențial și caracteristic pentru o serie de date, dacă acestea nu sunt prezentate într-un chip ordonat. Am mai văzut cum clasificarea datelor statistice într-o distribuție de frecvențe ne-a permis să punem în evidență în mod grafic anumite părți caracteristice ale fenomenului studiat. Varietatea tipurilor de distribuție descrise în capitolul precedent este o mărturie a acestui fapt.

Tendința aceasta de a căuta să prindem cu ajutorul unui șir de date măsurate sau observate în mod statistic caracterele esențiale ale unui fenomen constituie norma generală a metodei statistice, considerată ca procedeu de cercetare științifică. Numai astfel facem posibilă o discuție asupra fenomenului respectiv. Clasificarea și reprezentarea grafică a distribuției obținute prin clasificare nu sunt decât primii pași în această direcție.

Se caută apoi ca, făcând abstracție de forma distribuției, de mărimea intervalelor scării în raport cu care s-a făcut clasificarea, să se găsească modalitatea de a reprezenta printr-o singură valoare ordinul de mărime al datelor ce alcătuiesc o serie statistică.

Cum ordinul de mărime comun unui șir de valori este o noțiune care nu poate fi ușor și precis conturată, este foarte firesc ca, pentru acest scop în aparență unic, să fim conduși la utilizarea unor procedee diferite. Valorile la care ne conduce fiecare din aceste procedee sunt valori reprezentative ale seriei care a servit la calculul lor și poartă numele de valori tipice. Multă vreme statistica nu a avut în vedere decât considerarea acestor valori tipice. Din această cauză, am putea spune că azi tehnica valorilor tipice nu prezintă decât în mod cu totul excepțional puncte de întrebare științifice.

Vom menționa dintre valorile tipice pe cele mai uzitate.

42. *Media aritmetică.* Valoarea tipică cunoscută și întrebuințată în studiul fenomenelor colective încă din prima etapă istorică a metodei statistice este *media aritmetică.* Este prea cunoscut acest mod de a reprezenta un ansamblu de valori ca să mai fie nevoie de vreo introducere prealabilă în subiect.

Scanned with OKEN Scanner

Media aritmetică este bazată pe un principiu care ar trebui să ne călăuzească în orice comparaţie de date: raportul este singura valoare comparabilă.

În judeţul Dolj, pe o suprafaţă cultivată de 5 847 ha, viţa de vie a produs, în anul 1930, 156 644 hl, iar în judeţul Putna, în acelaşi an, s-au recoltat 1 089 927 hl dintr-o suprafaţă cultivată egală cu 16 970 ha. Comparaţia producţiei viţei de vie din aceste două judeţe este îngreunată de faptul că suprafeţele cultivate nu sunt aceleaşi. De aceea, printr-un procedeu de abstractizare, de altfel foarte natural, putem presupune că fiecare hectar a produs aceeaşi cantitate în interiorul aceluiaşi judeţ. Cantitatea aceasta poartă numele de producţie medie la hectar şi se obţine făcând raportul dintre producţia totală şi suprafaţa respectivă. Obţinem astfel o producţie medie la hectar de 26,9 hl pentru Dolj şi 64,2 hl pentru Putna. Acum, de-abia, putem face o comparaţie a producţiei viţei de vie din cele două judeţe.

Este de remarcat faptul că pentru calculul unei medii aritmetice nu este nevoie să cunoaştem toate valorile ce alcătuiesc seria, ci numai suma şi numărul lor. Acesta este unul din motivele pentru care media aritmetică este cea mai întrebuinţată, distribuţiile efective nefiind cunoscute decât în cazuri relativ rare. Pe un casier nu-l interesează distribuţia salariilor ce le are de plătit, ci mai ales totalul la care se urcă aceste salarii. Cunoscând acest total şi numărul salariaţilor, media aritmetică a salariilor respective este calculată imediat.

Media aritmetică mai poate fi interpretată şi în alt mod: ea este valoarea pe care, dacă ar avea-o toate elementele observate, totalitatea lor n-ar fi schimbată întru nimic. Ca o consecinţă a acestui fapt, media aritmetică poate fi considerată ca o valoare normală a măsurii unui fenomen. Cauzele specifice fiecărui caz în parte produc abateri de la această valoare. Dar atunci trebuie admis că efectele acestor cauze se compensează mutual. Această afirmaţie este însă discutabilă în majoritatea cazurilor. Cu toate acestea, nu este mai puţin adevărat că acest punct de vedere a fost cel care justifica în trecut întrebuinţarea mediei aritmetice. Se căuta să se dea astfel acestei valori o importanţă exagerată. Azi, fără ca importanţa ei să fie subestimată, media aritmetică nu este privită decât ca o bună valoare tipică.

43. *Calculul direct al mediei aritmetice*. Definiţia mediei aritmetice este mai uşor exprimată printr-o formulă matematică. Dacă x_1, x_2, x_3,..., x_p sunt valorile ce constituie ansamblul a cărui medie aritmetică vrem să aflăm, aceasta va fi dată de formula:

$$M = \frac{\sum x_i}{p}, \qquad (1)$$

unde semnul Σ este pus pentru a indica suma tuturor termenilor însemnaţi prin x.

Evident că dacă o valoare anumită este întâlnită de mai multe ori într-un ansamblu de date, ea trebuie să intervină în calculul mediei de atâtea ori de

Scanned with OKEN Scanner

câte ori a fost întâlnită, adică de atâtea ori câte unități reprezintă frecvența respectivă. Încât, dacă valorile x_1, x_2, ..., x_p nu se mai prezintă câte o singură dată ca în ipoteza de mai sus, ci de n_1, n_2,..., n_p ori respectiv, atunci media aritmetică va fi dată de formula:

$$M = \frac{\sum nx}{N}, \qquad (2)$$

unde N reprezintă totalitatea frecvențelor, $\sum_n$. Însemnând acum cu f_i frecvența relativă $\frac{n_i}{N}$, corespunzătoare frecvenței absolute n_i, formula precedentă devine:

$$M = \sum fx \qquad (3)$$

formă care are mai mult o importanță teoretică decât una practică, deoarece - deși mai simplă decât (2) - conduce pe un drum mai lung la valoarea lui M, necesitând calcularea prealabilă a frecvențelor relative.

Să aplicăm formula (2) la calculul mediei aritmetice a distribuției din tabelul 27. Mersul lucrărilor este arătat în tabelul 38.

TABELUL 38

Calcularea directă a mediei aritmetice

Numărul razelor pedunculului x	Frecvențele absolute n	Totalul razelor pe clase nx
(1)	(2)	(3)
5	1	5
6	5	30
7	9	63
8	22	176
9	38	342
10	62	620
11	61	671
12	29	348
13	14	182
14	4	56
15	4	60
	249	**2 553**

Tabelul 27

Împărţind acum totalul coloanei (3), care reprezintă numărătorul formulei (2), la totalul frecvenţelor 249, obţinem valoarea mediei aritmetice a distribuţiei considerate:

$$M = 10{,}25 \text{ raze.}$$

Dacă distribuţia nu mai este făcută în raport cu valori individuale ale variaţiei independente, cum a fost cazul precedent, ci în raport cu o scară de intervale, este absolut necesar ca pentru calcularea mediei aritmetice corespunzătoare să facem în prealabil anumite ipoteze.

În primul rând, trebuie remarcat un fapt:

Un număr oarecare de valori dintr-un ansamblu poate fi înlocuit cu o singură valoare, fără ca media întregului ansamblu să se schimbe, cu condiţia ca această valoare să fie media aritmetică a acestui grup de valori şi să i se atribuie ca frecvenţă numărul valorilor înlocuite.

Acest fapt, care pentru cei ce mânuiesc mai uşor instrumentul matematic constituie aproape o evidenţă, se poate uşor controla şi totodată interpreta cu ajutorul exemplului următor.

Să considerăm producţia de cărbuni a României (tabelul 23) şi să înlocuim producţiile din primii 4 ani prin media aritmetică a lor (tabelul 39). Din însuşi faptul menţionat mai sus, că media aritmetică a unui ansamblu nu depinde de valorile individuale, ci de suma lor, ceea ce decurge şi din formula (2), rezultă că media aritmetică a producţiei de cărbuni din România pentru perioada 1919-1930 este aceeaşi, fie că se utilizează datele din coloana (2), fie cele din coloana (3). Verificarea se face imediat: media aritmetică este egală şi într-un caz şi în altul cu 2 501 216 tone.

TABELUL 39

Producţia anuală de cărbuni a României

Anii	Producţia în tone	Serie transformată
(1)	(2)	(3)
1919	**1 559 330**	**1 766 953**
1920	**1 587 575**	**1 766 953**
1921	**1 804 687**	**1 766 953**
1922	**2 116 221**	**1 766 953**
1923	2 521 393	2 521 393
1924	2 776 371	2 776 371
1925	2 928 850	2 928 850
1926	3 053 553	3 053 553
1927	3 223 468	3 223 468
1928	3 027 240	3 027 240
1929	3 046 027	3 046 027
1930	2 369 882	2 369 882

Tabelul 23

Scanned with OKEN Scanner

Revenind acum la cazul unei distribuţii pe intervale, să presupunem că într-o distribuţie a unui grup de indivizi în raport cu înălţimea avem frecvenţa 4 corespunzătoare intervalului 180-181 cm. Despre indivizii încadraţi în acest interval nu ştim altceva decât că au înălţimile cuprinse între limitele indicate. Or, pentru calculul mediei aritmetice, este necesară cunoaşterea precisă a tuturor valorilor ce compun seria. Ignorându-le, le vom fixa după următoarele considerente:

În primul rând, este foarte natural să presupunem că taliile celor 4 indivizi consideraţi se repartizează în mod uniform în interiorul intervalului 180-181, adică vom avea câte o valoare cuprinsă în fiecare sfert al acestui interval: prima între 180-180,25, a doua între 180,25-180,50, ..., ultima între 180,75 - 181. În al doilea rând, e logic să atribuim fiecăreia din aceste talii o valoare egală cu mijlocul intervalului în care a fost repartizată, adică:

$$180{,}125\,,\quad 180{,}375\,,\quad 180{,}625\,,\quad 180{,}875^{1}.$$

Media acestor valori este egală cu 180,5. După principiul enunţat mai sus, urmează că, pentru calculul mediei aritmetice a întregii distribuţii, taliile din acest interval pot fi înlocuite prin altele patru egale cu 180,5, adică tocmai cu mijlocul intervalului considerat 180-181.

Acelaşi raţionament se poate face pentru oricare interval şi oricare altă distribuţie.

TABELUL 40

Calcularea directă a mediei aritmetice

Vârsta	Mijlocul intervalelor x	Frecvenţa n	Produsul nx
(1)	(2)	(3)	(4)
15 - 20	17,5	2	35
20 - 25	22,5	38	855
25 - 30	27,5	96	2 640
30 - 35	32,5	123	3 997,5
35 - 40	37,5	132	4 950
40 - 45	42,5	129	5 482,5
45 - 50	47,5	123	5 842,5
50 - 55	52,5	110	5 775
55 - 60	57,5	88	5 060
60 - 65	62,5	70	4 375
65 - 70	67,5	46	3 105

[1] *Aceasta, pentru că atunci când adevăratele valori nu sunt cunoscute, nu există nici un motiv care să ne facă să stricăm simetria acestei repartiţii.*

Scanned with OKEN Scanner

(continuare tab. 40)

(1)	(2)	(3)	(4)
70 - 75	72,5	26	1 885
75 - 80	77,5	12	930
80 - 85	82,5	4	330
85 - 90	87,5	1	87,5
		1 000	**45 350**

Tabelul 12

De aici se deduce următoarea regulă pentru calculul mediei aritmetice a unei distribuţii pe intervale: toate cazurile cuprinse într-un interval sunt considerate egale între ele şi egale cu mijlocul intervalului respectiv; apoi se trece la calculul mediei prin formula (2), aşa cum s-a făcut mai sus. Procedeul de calcul este pus în evidenţă prin exemplul din tabelul 40.

Totalul coloanei (4) ne dă imediat valoarea mediei aritmetice corespunzătoare distribuţiei considerate:

$$M = \frac{45\,350}{1000} = 45{,}35 \text{ani}\,.$$

44. *Scara intervalelor şi media aritmetică*. O chestiune care de foarte multe ori este neglijată atunci când se face calculul mediei aritmetice a unei distribuţii pe intervale este aceea a felului cum este fixat mijlocul intervalelor respective.

Am văzut, cu prilejul discuţiei asupra modului în care se face fixarea limitelor intervalelor scării de distribuţie (§ 18), că dacă se întrebuinţează forma (b), atunci intervalul exprimat prin limitele 1,0 - 1,9, de pildă, are ca limite efective 0,95 - 1,95, aceasta din cauza ordinului de aproximaţie adoptat pentru măsurarea elementelor seriei. Se vede însă că, oricum am calcula mijlocul intervalului, fie pornind de la limitele aproximative 1,0-1,9, fie pornind de la limitele efective 0,95 - 1,95, obţinem aceeaşi valoare egală cu 1,45.

Este o eroare ca în cazul unei astfel de scări să se considere mijlocul intervalului 1,0 - 1,9 egal cu 1,5. Eroarea aceasta este din ordinul celor care se numesc sistematice şi care se transmit în întregime asupra rezultatului operaţiei. Astfel, în cazul citat, cum eroarea făcută la fiecare interval este o majorare egală cu 0,05, media aritmetică va fi mărită cu această valoare.

Să considerăm cazul mediei aritmetice a distribuţiei satelor din judeţul Gorj (tabelul 41).

Media artimetică a acestei distribuţii este:

$$M = \frac{196133{,}5}{433} = 452{,}96\,.$$

Dacă mijloacele intervalelor ar fi fost fixate la valori terminate prin 50 în loc de 49,5, media ar fi fost mărită cu 0,5, adică ar fi fost egală cu 453,46.

Scanned with OKEN Scanner

TABELUL 41

Calculul mediei aritmetice

Mijlocul intervalelor	Frecvenţe	Produsul (1) x (2)
(1)	(2)	(3)
49,5	19	940,5
149,5	69	10 315,5
249,5	74	18 463
349,5	63	22 018,5
449,5	66	29 667
549,5	39	21 430,5
649,5	29	18 835,5
749,5	20	14 990
849,5	13	11 043,5
949,5	13	12 343,5
1 049,5	8	8 396
1 149,5	4	4 598
1 249,5	5	6 247,5
1 349,5	6	8 097
1 449,5	-	-
1 549,5	1	1 549,5
1 649,5	-	-
1 749,5	3	5 248,5
1 849,5	-	-
1 949,5	1	1 949,5
	433	**196 133,5**

Tabelul 15 B

45. *Stabilitatea mediei aritmetice*. Ultima medie aritmetică a fost calculată cu ajutorul distribuţiei din tabelul 15 B. Or, după cum ştim, distribuţia constituie prin ea însăşi o oarecare aproximaţie care este efectul clasificării pe intervale. Rezultă de aici că şi valoarea mediei astfel calculată nu va fi decât o aproximaţie a mediei adevărate.

În cazul considerat putem face chiar o comparaţie între aceste două valori. Într-adevăr, găsim în *Indicatorul statistic* citat, p. 174, că totalul populaţiei rurale a judeţului Gorj este egal cu 195 155 locuitori. Media aritmetică adevărată va fi deci:

$$M' = \frac{195\,155}{433} = 450{,}70\,.$$

Aceleaşi comune urbane au fost distribuite şi în chipul înfăţişat de tabelul 28. Media aritmetică calculată cu ajutorul acestei ultime distribuţii este:

Scanned with OKEN Scanner

$$M'' = \frac{204\,183{,}5}{433} = 471{,}56\,.$$

Din compararea valorilor *M*, *M'*, *M''* vedem că, deşi aceste valori au fost calculate plecând de la forme diferite de prezentare a aceluiaşi material, ele nu diferă prea mult între ele. Singura care se distanţează mai mult este valoarea *M''*. Lucrul se explică prin faptul că distribuţia după care s-a calculat această valoare este făcută pe o scară cu intervale prea mari.

De aici, două puncte de relevat:

1. Media aritmetică este o valoare tipică *stabilă*, cu alte cuvinte nu este prea mult influenţată de schimbări făcute în ansamblul valorilor cărora le corespunde, cu condiţia ca aceste schimbări să nu fie fundamentale.
2. Distribuţiile făcute în raport cu o scară de intervale prea mari trebuie evitate, deoarece nu permit o evaluare satisfăcătoare a mediei. Dar, în acelaşi timp, nu trebuie să căutăm să îngreunăm forma distribuţiei, adoptând intervale prea mici, căci după cum am văzut, rămânând la intervale de mărime optimă, obţinem media aritmetică cu aproximaţie suficientă.

46. *Compunerea mediilor aritmetice*. Media aritmetică se mai bucură şi de alte proprietăţi, între care cea mai importantă este aceea că admite o *teoremă de adiţie*. Se înţelege prin acest lucru posibilitatea ce o avem de a calcula media aritmetică a unui ansamblu compus din două grupuri de valori, cunoscând numai mediile aritmetice ale celor două grupuri, precum şi numărul de elemente ce le constituie. Astfel, dacă m_1, n_1; m_2, n_2 sunt respectiv mediile aritmetice şi numărul elementelor celor două grupuri, media ansamblului lor va fi:

$$M = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}. \qquad (4)$$

Media aritmetică a numărului de morţi din România pentru anii 1901-1905 este egală cu 160 861 decese, iar pentru anii 1906-1915, cu 179 246 (tabelul 24). Media aritmetică pentru anii 1901-1915 va fi, după formula (4):

$$M = \frac{5 \times 160\,861 + 10 \times 179\,246}{5 + 10} = 173\,118\,.$$

De altfel, formula (4) este valabilă pentru un număr oarecare de grupuri:

$$M = \frac{n_1 m_1 + n_2 m_2 + \ldots + n_s m_s}{n_1 + n_2 + \ldots + n_s} = \frac{\sum nm}{\sum n}. \qquad (5)$$

Scanned with OKEN Scanner

Un caz particular al acestei formule, foarte des întâlnit în practică, este acela când grupurile au același număr de elemente constitutive și când formula devine:

$$M = \frac{m_1 + m_2 + .. + m_s}{s}. \qquad (5 \text{ bis})$$

Este de remarcat analogia dintre formula (5) și (3): *obținem media aritmetică a unui ansamblu format din mai multe grupuri, făcând media mediilor aritmetice ale acestor grupuri și atribuind fiecăreia din ele o frecvență egală cu numărul elementelor respective.*

De altfel, pe acest fapt se bazează și operațiunea de înlocuire a elementelor din intervalele unei distribuții prin media presupusă a lor, atunci când se calculează media aritmetică corespunzătoare acesteia.

47. *Calculul mediei aritmetice prin deplasarea originii.* Procedeul întrebuințat mai sus este calea directă pentru calcularea mediei aritmetice. Oricine își poate da seama că acest procedeu necesită calcule care nu sunt din cele mai scurte. Calculele pot fi mult prescurtate dacă se urmează așa-numitul procedeu de calcul prin *schimbarea originii.* Acest procedeu este bazat pe următorul fapt: *dacă se măresc sau se micșorează toate valorile unui ansamblu cu aceeași cantitate, media aritmetică va fi mărită sau micșorată la fel.*

Se măsoară înălțimile, de la pământ până la creștetul capului, ale unui grup de indivizi; se calculează media aritmetică a grupului. Printr-un mijloc oarecare se ridică întregul grup de la pământ cu o distanță oarecare; media aritmetică a înălțimilor astfel obținute va fi, evident, mărită cu aceeași cantitate față de precedenta. Se vede acum de ce se întrebuințează termenul de schimbarea originii: înălțimile, în loc să fie măsurate de la nivelul normal, sunt considerate ca începând de la un alt punct.

Dacă α_1, α_2,... α_i sunt valorile corespunzătoare ansamblului considerat, măsurate de la noua origine *a*, atunci vom obține media aritmetică adevărată, dacă la media aritmetică a valorilor α adăugăm valoarea *a*:

$$M = a + \frac{\sum n\alpha}{\sum n} \qquad (6)$$

relație care se poate exprima mai sugestiv sub forma:

$$M = a + M(a), \qquad (6 \text{ bis})$$

unde *M (a)* înseamnă media aritmetică calculată luând valoarea *a* ca origine.

Valorile α se numesc cu un termen tehnic abaterile în raport cu *a* și se obțin evident prin diferența:

$$\alpha = x - a.$$

Scanned with OKEN Scanner

Pentru ilustrarea acestui nou procedeu de calcul, să considerăm din nou distribuția din tabelul 27 și să adoptăm ca origine valoarea 10, ceea ce înseamnă că toate valorile vor fi *diminuate cu această cantitate*. Metoda de calcul reiese din tabelul 42.

TABELUL 42

Calcularea mediei aritmetice prin schimbarea originii

Numărul razelor pedunculului x	Frecvențe n	Abateri în raport cu 10 α	Produsul $n\alpha$
(1)	(2)	(3)	(4)
5	1	- 5	- 5
6	5	- 4	- 20
7	9	- 3	- 27
8	22	- 2	- 44
9	38	- 1	- 38
10	62	0	**- 134**
11	61	+ 1	+ 61
12	29	+ 2	+ 58
13	14	+ 3	+ 42
14	4	+ 4	+ 16
15	4	+ 5	+ 20
	249	-	**+ 197**

Tabelul 27

Totalul algebric al coloanei (4) va fi + 197 - 134 = + 63 și deci:

$$M(a) = \frac{+63}{249} = 0{,}25$$

și, prin urmare:

$$M = 10 + 0{,}25 = 10{,}25$$

valoarea obținută și pe cale directă (§ 43).

Valoarea aleasă ca origine este arbitrară, însă noi trebuie să profităm de acest lucru pentru ca să facem calculele cât mai simple posibil. În general, este preferabil ca valoarea aceasta să fie astfel aleasă, încât să împartă scara distribuției în două părți aproape egale. Numai astfel evităm abaterile prea mari.

Procedeul de calcul al mediei aritmetice poate să fie simplificat și mai mult. Abaterile α pot fi măsurate în unități egale cu intervalele. Să considerăm distribuția din tabelul 31. Mărimea intervalelor acestei distribuții este egală cu 5 livre. Să adoptăm ca origine valoarea 135 (tabelul 43).

Scanned with OKEN Scanner

Se deduce:

$$M(a) = \frac{-782 + 147}{416} \text{ int } ervale = 1{,}526 \text{ intervale}$$
$$= -1{,}526 \times 5 = -7{,}630 \text{ livre}$$

deci:

$$M = 135 - 7{,}630 = 127{,}37 \text{ livre.}$$

TABELUL 43

Calculul mediei aritmetice prin schimbarea originii

Greutatea	Mijlocul intervalelor x	Frecvențe n	Abateri în intervale α	Produsul $n\alpha$
(1)	(2)	(3)	(4)	(5)
97,5 –	100	3	- 7	- 21
102,5 –	105	6	- 6	- 36
107,5 –	110	23	- 5	- 115
112,5 –	115	48	- 4	- 192
117,5 –	120	56	- 3	- 168
122,5 –	125	86	- 2	- 172
127,5 –	130	78	- 1	- 78
132,5 –	135	45	0	**- 782**
137,5 –	140	30	+ 1	+ 30
142,5 –	145	22	+ 2	+ 44
147,5 –	150	11	+ 3	+ 33
152,5 –	155	3	+ 4	+ 12
157,5 –	160	3	+ 5	+ 15
162,5 –	165	1	+ 6	+ 6
167,5 –	170	1	+ 7	+ 7
		416		**+ 147**

Tabelul 31

Se vede din tabelul 43 că dacă n-am fi măsurat abaterile în intervale, coloana (4), și deci și coloana (5), ar fi conținut numere mai mari și astfel calculele ar fi necesitat mai mult timp.

Procedeul de mai sus poate fi sintetizat prin următoarea formulă:

$$M = a + \frac{\sum n\alpha}{\sum n} \cdot d, \qquad (7)$$

unde d este mărimea comună a intervalelor distribuției, iar α abaterile măsurate în intervale.

Scanned with OKEN Scanner

48. *Abaterile în raport cu media aritmetică.* Să presupunem acum că originea aleasă este egală chiar cu media aritmetică a seriei considerate. În acest caz, relaţia (6) devine:

$$M = M + \frac{\sum n\xi}{\sum n},$$

de unde:

$$\frac{\sum n\xi}{\sum n} = 0 \quad sau \quad \sum n\xi = 0, \tag{8}$$

unde am însemnat prin ξ_i abaterea valorii x_i în raport cu media aritmetică:

$$\xi_i = x_i - M.$$

Relaţiile (8) constituie una din proprietăţile caracteristice ale mediei aritmetice:

Media aritmetică (suma) abaterilor valorilor unei serii în raport cu media aritmetică a acesteia este egală cu zero. Media aritmetică este singura valoare bucurându-se de această proprietate.

Vom avea mai târziu ocazia să utilizăm relaţiile (8); deocamdată ne oprim doar asupra unei verificări numerice a acestei proprietăţi (tabelul 44).

TABELUL 44

Abaterile în raport cu media aritmetică

Vârsta x	Abateri în raport cu media 45,35 δ	Frecvenţe n	Produsul $n\delta$
(1)	(2)	(3)	(4)
17,5	- 27,85	2	- 55,70
22,5	- 22,85	38	- 868,30
27,5	- 17,85	96	- 1 713,60
32,5	- 12,85	123	- 1 580,55
37,5	- 7,85	132	- 1 036,20
42,5	- 2,85	129	- 367,65
47,5	+ 2,15	123	+ 264,45
52,5	+ 7,15	110	+ 786,50
57,5	+ 12,15	88	+ 1 069,20
62,5	+ 17,15	70	+ 1 200,50

Scanned with OKEN Scanner

(continuare tab. 44)

(1)	(2)	(3)	(4)
67,5	+ 22,15	46	+ 1 018,90
72,5	+ 27,15	26	+ 705,90
77,5	+ 32,15	12	+ 385,80
82,5	+ 37,15	4	+ 148,60
87,5	+ 42,15	1	+ 42,15
		1 000	**- 5 622,00**
			+ 5 622,00

Tabelul 40

$$\sum n\xi = -5\,622 + 5\,622 = 0.$$

49. *Valoarea mediană*. Media aritmetică considerată ca valoare tipică reprezintă, după cum am văzut mai sus, un nivel mijlociu pe care l-ar avea toate elementele din serie dacă, păstrându-se valoarea lor globală, acestea ar fi egale între ele. Tot pe baza acestui principiu, de a reprezenta o serie printr-o valoare *mijlocie*, se întrebuințează o altă valoare tipică, numită *valoare mediană.*

Mediana este egală cu valoarea elementului care, toate elementele seriei fiind aranjate în ordine de mărime, ar împărți acest șir în două grupe egale ca număr. Este deci ceea ce am putea numi într-adevăr o valoare mijlocie.

Valoarea medianei este evident egală cu cel de al $(n + 1)$-lea element în ordinea de mărime, dacă seria are $2n + 1$ elemente. Acest element lasă deasupra valorii sale și sub ea câte n valori. În cazul când seria este compusă din $2n$ elemente, nu există element care să fie central. S-a făcut convenția ca în acest caz să se adopte pentru valoarea medianei media aritmetică a elementelor de rangul n și $n + 1$.

Astfel, dacă se consideră seria statistică din tabelul 24, aranjând elementele constitutive ale ei în ordinea lor descrescândă, ca în tabelul 45, vom avea pentru mediana corespunzătoare valoarea elementului de rangul 8, adică 25,2 decese la mia de locuitori.

TABELUL 45

Calculul medianei

Rangul	Anul	Numărul deceselor la 1 000 locuitori
(1)	(2)	(3)
1	1912	22,9
2	1914	23,5
3	1906	23,9
4	1904	24,4
5	1915	24,5

Scanned with OKEN Scanner

(continuare tab. 45)

(1)	(2)	(3)
6	1905	24,7
7	1910	24,8
8	**1903**	**25,2**
9	1911	25,3
10	1913	26,1
11	1907	26,3
12	1901	26,5
13	1908	27,4
14	1909	27,4
15	1902	28,1

Tabelul 24

Trecând acum la seria din tabelul 46, care conține un număr par de valori, vom avea pentru valoarea mediană media aritmetică a valorilor de rangul 5 și 6:

$$M_e = \frac{282342 + 299870}{2} = 291106 .$$

Noțiunea de valoare mediană poate fi generalizată. Putem adopta ca valoare mediană media aritmetică a unui număr mai mare de termeni centrali. Vom lua, după cum am văzut și în cazul medianei simple, un număr par de valori când seria cuprinde $2n$ elemente și un număr impar de valori în cazul contrar. Elementele care ne vor servi de bază pentru calculul medianei generalizate vor trebui să fie simetric ordonate în raport cu termenul sau termenii care ne conduc la mediana simplă.

TABELUL 46

Calculul medianei

Rangul	Anul	Numărul născuților
1	1906	262 438
2	1908	272 850
3	1910	273 106
4	1907	274 487
5	**1909**	**282 342**
6	**1911**	**299 870**
7	1913	309 625
8	1912	314 090
9	1915	319 544
10	1914	327 345

Tabelul 25

Scanned with OKEN Scanner

În cazul seriei din tabelul 45, cum elementul central are rangul 8, vom putea adopta pentru mediana generalizată media aritmetică a elementelor de rangul 7, 8, 9, adică 25,1, sau a elementelor de rangul 6, 7, 8, 9, 10, adică 25,22 etc. Pentru seria din tabelul 46, putem lua, de pildă, media aritmetică a elementelor de rangul 4, 5, 6, 7, a cărei valoare este de 291 581.

Mediana generalizată poate cuprinde un număr din ce în ce mai mare de termeni, iar la limită, pe toți. În acest din urmă caz, ea se confundă cu media aritmetică a seriei. Din această cauză, mediana generalizată poate să fie cu folos întrebuințată în locul mediei aritmetice, în cazul seriilor care cuprind valori excepționale, fie prea mari, fie prea mici, și care ar influența simțitor valoarea acesteia din urmă (§ 53).

50. *Calculul valorii mediane pentru distribuțiile heterograde.* Am văzut care este procedeul prin care putem determina valoarea medianei, în cazul când cunoaștem valorile individuale ale seriei statistice. Să trecem acum la cazul unei distribuții de frecvențe și, în acest scop, să considerăm distribuția unui grup de flori de mac de câmp, în raport cu numărul stigmatelor (tabelul 47).

TABELUL 47

Distribuția florilor de mac, în raport cu numărul stigmatelor

Numărul stigmatelor	Frecvențe	Frecvențe cumulate
7	3	3
8	**9**	**12(3 + 9)**
9	4	16(12 + 4)
10	3	19(16 + 3)
11	1	20(19 + 1)

Extras din E. Coquidé, ***Amélioration des plantes cultivées et du bétail****, Paris, Baillière, 1920, p.101.*

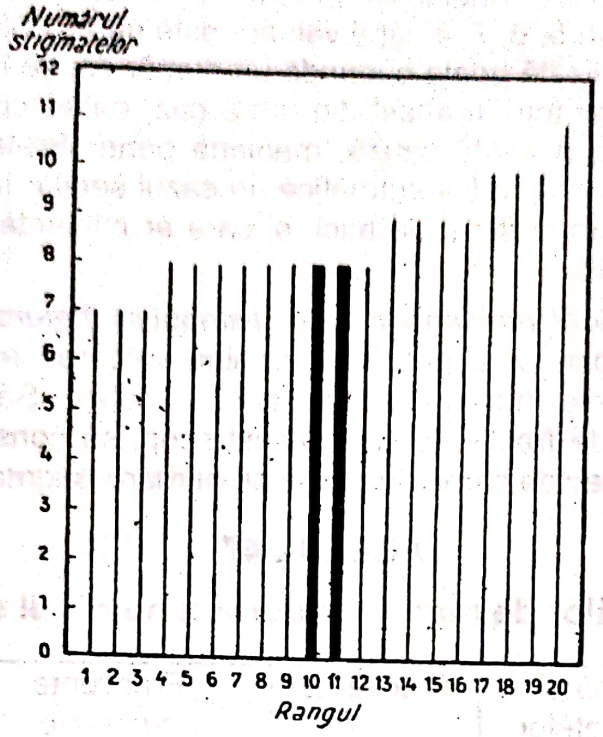
Valoarea mediană va fi egală, după însăși definiția acesteia, cu media aritmetică a elementelor de rangul 10 și 11. Pentru a afla care sunt aceste elemente, vom deduce din distribuția considerată *distribuția cumulativă* a aceluiași fenomen. Am întâlnit mai sus forma cumulativă a unei serii statistice în timp (§13). Forma cumulativă a unei distribuții de frecvențe se va obține printr-un procedeu analog: se cumulează frecvențele începând cu cea corespunzătoare unui interval marginal - de obicei cel inferior - și continuând în ordinea lor naturală.

Se vede din tabelul 47 că 12 flori au avut mai puțin de 9 stigmate și că 3 au avut mai puțin de 8. Prin urmare, elementele de rangul 10 și 11 sunt cuprinse în categoria exemplarelor cu 8 stigmate și deci, pentru distribuția considerată:

Me = 8 stigmate.

Scanned with OKEN Scanner

Pentru a face acest procedeu mai uşor de urmărit, am redat printr-un grafic (graficul 25) ordonarea elementelor ce-au format distribuţia considerată, punând în evidenţă poziţia elementului median.



Tabelul 47

Graficul 25. Poziţia valorilor care conduc la valoarea mediană.

51. Chestiunea calculului medianei în cazul exemplului precedent nu a întâmpinat nici o dificultate, din cauză că distribuţia a fost făcută în raport cu valori individuale ale variabilei independente. Când însă distribuţia este făcută în raport cu o scară de intervale, calculul medianei se sprijină pe ipoteza repartiţiei uniforme a elementelor dintr-un interval, ipoteză de care am avut nevoie şi pentru calcularea mediei aritmetice.

Să revenim la distribuţia din tabelul 51. Vom calcula, în primul rând, forma cumulativă corespunzătoare acestei distribuţii (tabelul 48).

Pentru calculul valorii mediane a acestei distribuţii trebuie să cunoaştem elementele de rangul 208 şi 209. Acestea sunt cuprinse în intervalul 122,5 - 127,5, care poartă numele de *interval median*. Făcând ipoteza că elementele sunt distribuite în mod uniform în acest interval, îl vom împărţi în 86 de părţi egale cu $\frac{5}{86}$ livre. Cum primul element din acest interval are rangul 137, va trebui, pentru a afla elementul de rangul 208, să căutăm pe acela care în interiorul intervalului considerat are rangul 208 - 136 = 72. Şi cum fiecare element din acest interval diferă de precedentul prin valoarea $\frac{5}{86}$ livre, urmează că elementele căutate vor fi egale cu:

Elementul de rangul 208 = $122,5 + \frac{5 \times 72}{86} = 126,686$

Elementul de rangul 209 = $122,5 + \frac{5 \times 73}{86} = 126,744$

Deci valoarea medianei va fi:

$Me = 126,715$ livre.

TABELUL 48

Calculul formei cumulative a unei distribuții heterograde

Greutatea în livre	Frecvențe	Frecvențe cumulate
97,5 -	3	3
102,5 -	6	9
107,5 -	23	32
112,5 -	48	80
117,5 -	56	136
122,5 -	**86**	**222**
127,5 -	78	300
132,5 -	45	345
137,5 -	30	375
142,5 -	22	397
147,5 -	11	408
152,5 -	3	411
157,5 -	3	414
162,5 -	1	415
167,5 -	1	416

Tabelul 31

Observăm în legătură cu această valoare două fapte:

1. Valoarea medianei ar fi fost aceeași dacă, printr-o abstractizare, naturală în domeniul valorilor continue, am fi atribuit acestei valori mărimea elementului de rangul 208,5 - media aritmetică a rangurilor 208 și 209:

$$Me = 122,5 + \frac{5 \times 72,5}{86} = 126,715 \text{ livre}.$$

2. Astfel obținută valoarea medianei, aceasta este supraestimată cu o jumătate din $\frac{5}{86}$ livre.

Într-adevăr, primul element din intervalul 122,5 - 127,5 nu este egal cu $122,5 + \frac{5}{86}$, ci cuprins între 122,5 şi $122,5 + \frac{5}{86}$: îi vom atribui, conform ipotezei făcute, valoarea $122,5 + \frac{1}{2} \cdot \frac{5}{86}$. Pornind de aici, fiecare din elementele următoare se obţine prin adăugarea succesivă a valorii $\frac{5}{86}$. Aşadar, o valoare judicioasă a medianei va fi:

$$Me = 122,5 + \frac{5 \times 72}{86} = 126,686\ ^{1}.$$

Aceste considerente ne conduc la formularea următoarei reguli:

Valoarea mediană a unei distribuţii pe intervale este dată de elementul de rangul $\frac{N}{2}$, *unde N este numărul total al frecvenţelor, indiferent dacă N este par sau impar.*

Regula aceasta conduce la următoarea formulă:

$$Me = x_i + \frac{N - 2N_{i-1}}{2n_i} \cdot d, \qquad (9)$$

unde x_i este limita inferioară a intervalului median, n_i, frecvenţa corespunzătoare acestui interval, d, mărimea lui, şi N_{i-1}, frecvenţa cumulată a intervalului imediat precedent.

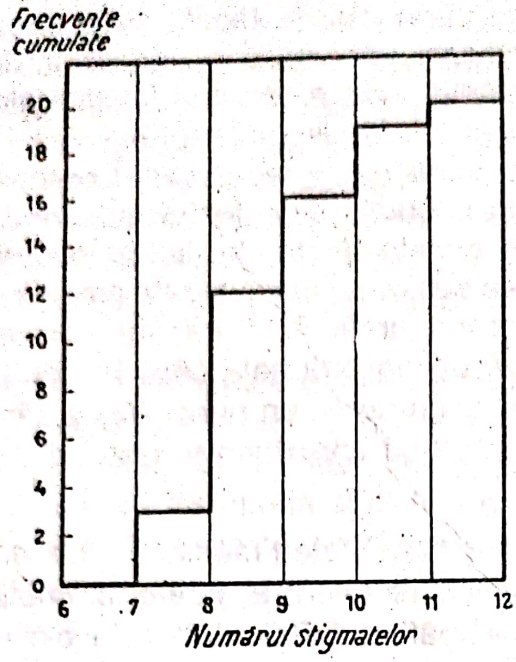
52. *Ogiva. Evaluarea grafică a medianei.* Dacă valoarea medianei nu trebuie obţinută cu aproximaţie prea bună, atunci, pentru aflarea ei, se poate întrebuinţa şi un procedeu grafic. Să considerăm distribuţia sub formă cumulativă din tabelul 47. Această distribuţie se poate reprezenta grafic după aceleaşi norme care ne-au călăuzit pentru reprezentarea distribuţiei simple. Astfel, se aşază pe axa absciselor valorile variabilei independente, în dreptul cărora se ridică în ordonate, de data aceasta, valori proporţionale *cu frecvenţele cumulate* corespunzătoare. Extremităţile acestor ordonate ne vor conduce la trasarea unei linii care poartă numele de *ogivă*.

Ogiva se completează în chip diferit, după cum distribuţia respectivă este continuă sau nu.

În cazul distribuţiei din tabelul 47, avem de-a face cu o distribuţie de variabilă discontinuă. Ogiva va avea forma înfăţişată de graficul 26: o linie în formă de scară. Şi este natural să fie astfel. Între valorile 9 şi 10, frecvenţa cumulată trebuie să rămână constantă, deoarece nici un exemplar nu poate avea un număr fracţionar de stigmate.

---

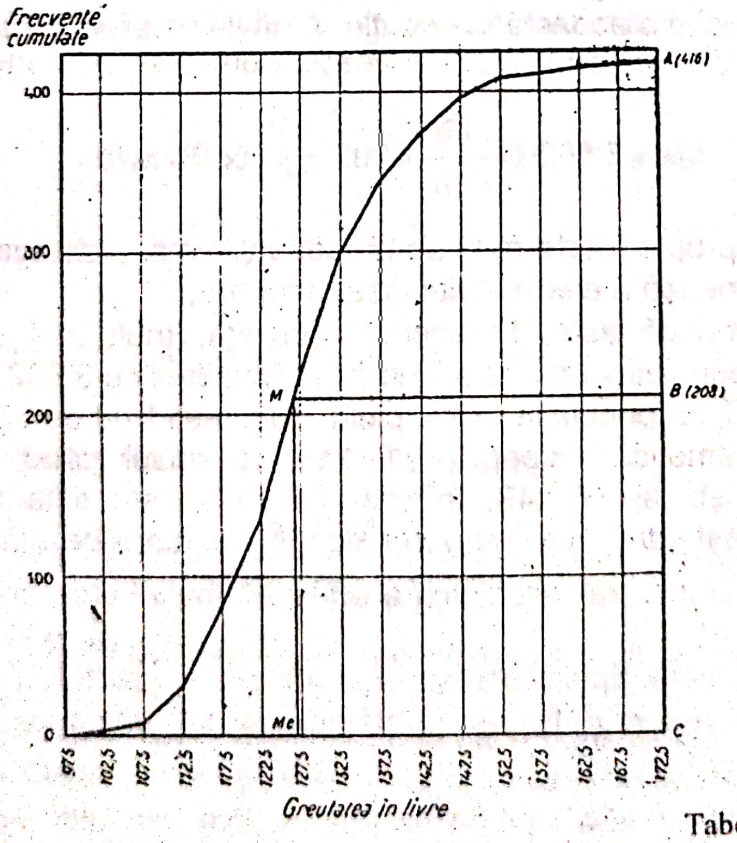
[1] *Diferenţa între această valoare şi precedenta (126,715) provine din faptul că se trece din domeniul discontinuu în domeniul continuu.*



Tabelul 47

Graficul 26. Ogiva unei distribuții de variabilă discontinuă.

Dacă însă distribuția este continuă - cum este cazul celei din tabelul 48 - ogiva ia forma unei linii poligonale (graficul 27).



Tabelul 48

Graficul 27. Determinarea grafică a medianei, prin intermediul ogivei.

Scanned with OKEN Scanner

Forma aceasta conţine în ea ipoteza făcută asupra repartiţiei uniforme a elementelor în interiorul aceluiaşi interval, ceea ce se traduce pe grafic prin faptul că extremităţile ordonatelor care reprezintă frecvenţele cumulate a două intervale consecutive sunt unite printr-o linie dreaptă.

Ceea ce mai este de observat în legătură cu construirea ogivei corespunzătoare unei distribuţii pe intervale este faptul că ordonatele reprezentând frecvenţele cumulate trebuie trasate în dreptul limitei superioare a intervalului respectiv, deoarece cumularea cuprinde absolut toate cazurile dintr-un interval.

Dacă pe un astfel de grafic (graficul 27) trasăm o dreaptă paralelă cu axa absciselor şi care să împartă ultima ordonată (*AB*) în două părţi egale (*AC* = *CB*), obţinem prin intersecţia ei cu ogiva un punct (*M*) a cărui abscisă (*Me*) ne dă valoarea medianei respective cu aproximaţie destul de bună în cazul când desenul este îngrijit făcut.

53. *Stabilitatea medianei.* Stabilitatea medianei - condiţie esenţială pentru o valoare tipică - este şi mai mare decât a mediei aritmetice. Explicaţia este simplă: un element a cărui valoare este excepţională din punct de vedere al mărimii influenţează totalul valorilor care ne dă valoarea mediei aritmetice; acelaşi element nu influenţează mediana decât prin rangul lui. Or, rangul unui element poate rămâne acelaşi, chiar dacă valoarea lui se schimbă simţitor.

Să considerăm forma cumulativă a distribuţiei din tabelul 13 B. Se întâmplă în cazul de faţă că valoarea mediană este limita dintre două intervale. Ea va fi, prin urmare, egală cu 9 999 locuitori. Să eliminăm, de exemplu, municipiul Bucureşti din această distribuţie. Atunci valoarea mediană va fi, după formula (9):

$$Me = 5\,000 + \frac{113}{114} \cdot 4\,999 = 9\,955 \text{ locuitori},$$

valoare care se apropie foarte mult de precedenta, mai ales dacă ţinem seama de importanţa numerică a elementului lăsat deoparte.

Media aritmetică este, în această privinţă, mult mai sensibilă decât mediana. Totalul populaţiei urbane a României fiind egal cu 3 632 178 (*Indicatorul statistic*, p. XXXVI), iar populaţia municipiului Bucureşti fiind de 631 288, obţinem pentru media aritmetică valoarea 21 117, în cazul când considerăm şi municipiul Bucureşti, şi 17 549, în cazul când acesta este lăsat deoparte. Diferenţa între aceste două valori nu mai poate fi desconsiderată.

54. *Modul.* O altă valoare tipică a unei distribuţii heterograde este *modul sau dominanta*[1].

Această valoare tipică corespunde tendinţei spiritului omenesc de a extinde calitatea cea mai frecvent întâlnită la tot ansamblul de elemente

[1] *Termenul englezesc* mode, *introdus de prof. K. Pearson, care s-a ocupat mai de aproape de această valoare tipică, a fost tradus în franceză prin acela de* dominante *(prof. Lucien March).*

Scanned with OKEN Scanner

cercetate. Dacă însă această tendință e dăunătoare judecății lucrurilor, valoarea tipică la care a dat naștere este una dintre cele mai utile. Dominanta este valoarea căreia îi corespunde frecvența maximă. Pe grafic, valoarea dominantă este dată de vârful poligonului frecvențelor.

În cazul când poligonul de frecvențe nu prezintă decât un vârf, un maximum, valoarea dominantă nu se calculează: ea se citește. Astfel, pentru distribuția de flori de *Pimpinella saxifraga*, dominanta este egală cu 10 (tabelul 27).

Dacă distribuția este continuă, atunci se ia pentru valoarea dominantă mijlocul intervalului de frecvență maximă. Astfel, modul distribuției din tabelul 31 este egal cu 125.

Cu toate acestea, modul este o valoare care depinde foarte mult de felul cum este alcătuită scara distribuției. O schimbare neînsemnată a acesteia și modul s-a deplasat (tabelul 50). De fapt, nu se poate vorbi despre mod atâta timp cât nu s-a calculat pe distribuția statistică o curbă de frecvențe teoretică. Această operație nu este însă din cele mai simple.

Se poate remedia în parte acest neajuns în mai multe feluri.

În primul rând, se calculează modul prin formule care să facă să intre în joc și frecvențele claselor vecine celei de frecvență maximă. O formulă clasică în acest scop este următoarea:

$$Mo = x_i + \frac{n_{i+1} - n_{i-1}}{n_{i+1} + n_{i-1}} \cdot \frac{d}{2}, \qquad (10)$$

unde x_i este mijlocul intervalului de frecvență maximă, n_{i+1} și n_{i-1}, frecvențele claselor adiacente. După formula aceasta, modul distribuției din tabelul 31 este:

$$M_0 = 125 + \frac{78 - 56}{78 + 56} \times \frac{5}{2} = 125{,}410 \text{ livre}.$$

O altă formulă în uz este:

$$M_0 = x_i + \frac{n_{i+1} - n_{i-1}}{2n_i - n_{i+1} - n_{i-1}} \cdot \frac{d}{2} \qquad (11)$$

care ne dă, pentru distribuția considerată, valoarea:

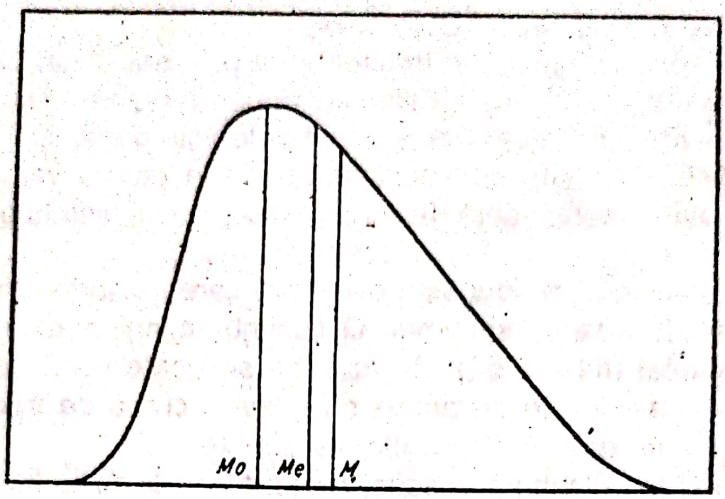
$$M_0 = 125 + \frac{78 - 56}{172 - 56 - 78} \times \frac{5}{2} = 126{,}447 \text{ livre}.$$

Formula (11) ne permite să ne apropiem mai mult de valoarea teoretică a modului decât (10).

Mai există un procedeu pentru aflarea modului, dar care nu se poate aplica decât în cazul distribuțiilor ușor asimetrice. Pentru astfel de distribuții,

Scanned with OKEN Scanner

valoarea mediei aritmetice, figurată pe grafic, se găsește de partea cozii lungi a acesteia, iar mediana la 2/3 între ea și mod (graficul 28)[1].



Graficul 28. Distribuție ușor asimetrică.

Deci

$$Mo = M - 3(M - Me) \qquad (12)$$

de unde, pentru cazul distribuției din tabelul 31, care este într-adevăr ușor asimetrică, obținem:

$$Mo = 127{,}37 - 3(127{,}37 - 126{,}69) = 125{,}33.$$

TABELUL 49

Comparație între valoarea modului, calculată după (11), și valoarea adevărată, pentru cazul câtorva distribuții de înălțimi barometrice

Stațiunea	Media aritmetică	Mediana	Modul crud	Modul adevărat
Southampton	29 981	30 000	30 038	30 039
Londonderry	29 891	29 915	29 963	29 960
Carmarthen	29 952	29 974	30 018	30 013
Glasgow	29 886	29 906	29 946	29 967
Dundee	29 870	29 890	29 930	29 951

*După G. U. Yule, **op. cit.**, p. 122.*

[1] *Mediana fiind valoarea pentru care frecvențele elementelor superioare și inferioare ei sunt egale, ordonata corespunzând acestei valori va împărți în două părți egale suprafața cuprinsă între poligonul de frecvență (histogramă sau curba frecvențelor) și axa absciselor.*

Scanned with OKEN Scanner

Această metodă de calcul a modului este recomandată de majoritatea autorilor. Tabelul 49 ne arată într-adevăr că acest procedeu ne permite o evaluare a adevăratului mod, cum nu se poate mai bine. Dar nu este mai puțin adevărat că această concordanță de valori este datorată, aici, și regularității fenomenului considerat.

55. *Distribuții plurimodale*. Acestea sunt procedeele de calculare a modului în cazul când distribuția prezintă o singură clasă de frecvență maximă. Chestiunea se complică în cazul când distribuția prezintă mai multe intervale modale, fapt care se traduce pe grafic prin prezența mai multor vârfuri ale poligonului de frecvență. Astfel de distribuții poartă numele de *distribuții plurimodale*; acestea au un mod principal și mai multe alte moduri secundare.

Distribuția din tabelul 50 constituie un exemplu de distribuție plurimodală (graficul 29). Frecvențele tipărite în caractere groase sunt cele care ne determină modurile acestei distribuții. Metoda care se urmează pentru calculul modului principal este cunoscută sub numele de *metoda grupărilor succesive*. Această metodă este bazată pe o observație care am făcut-o la timp: cu cât intervalele sunt mai mari, cu atât forma distribuției începe să devină mai regulată, asperitățile graficului respectiv dispărând treptat.

Așadar, metoda grupărilor va consta din gruparea intervalelor distribuției, la început două câte două, apoi trei câte trei și așa mai departe, până când se ajunge la o formă *unimodală*.

Se vede din tabelul 50 cum numărul modurilor descrește pe măsură ce intervalele se măresc. Ultima coloană nu mai prezintă decât un singur mod; acesta este dat de intervalul 1,05 - 1,54, căruia îi corespunde frecvența 2012. Modul principal al distribuției considerate, calculat după principiul de mai sus, va fi, prin urmare, egal cu 1,295.

Graficul 29 ne pune în evidență mobilitatea poziției modului pe scara distribuției; se vede cum intervalul modal se deplasează când trecem de la un fel de grupare la altul și cum alura generală a poligonului frecvențelor devine în același timp mai regulată.

Scanned with OKEN Scanner

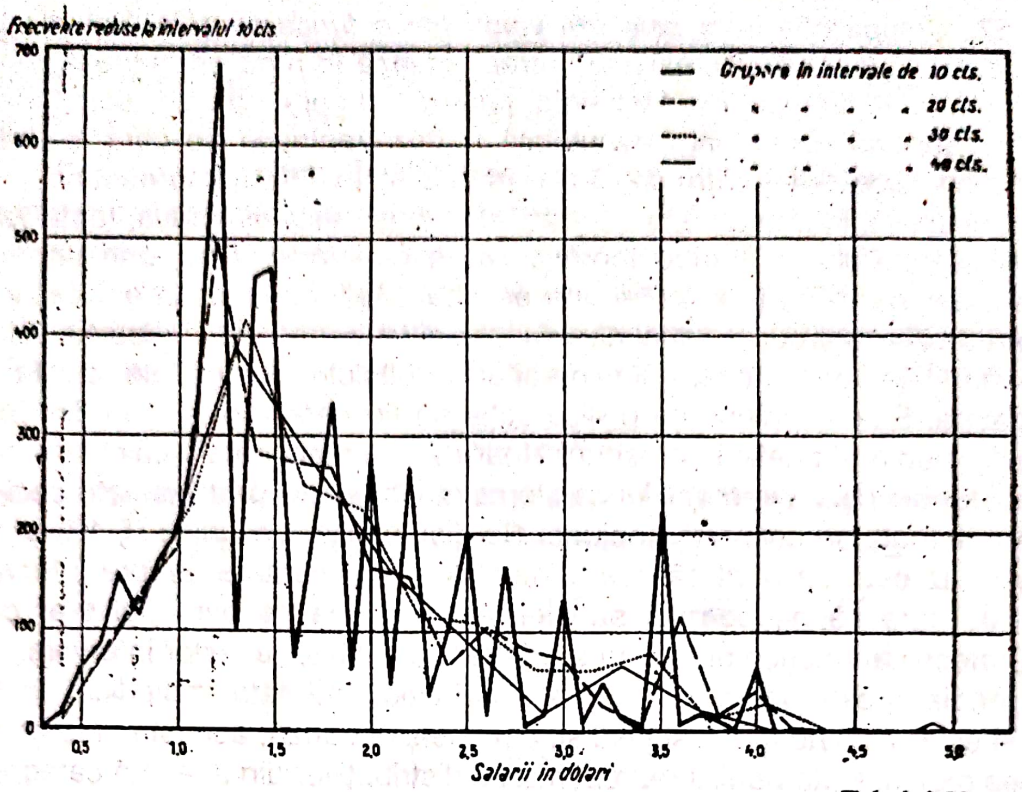
TABELUL 50

Distribuție de salarii, după Aldrich Report

Salarii în dolari	FRECVENȚE				
	Grupe de 10 cts.	Grupe de 20 cts.	Grupe de 30 cts.	Grupe de 40 cts.	Grupe de 50 cts.
.25- .34	1				
.35- .44	15	16			
.45- .54	59	74	75		
.55- .64	85	144	159	160	
.65- .74	**157**	242	355	316	317
.75- .84	113	270	301	414	725
.85- .94	169	282	439	524	
.95- 1.04	201	370	674	640	
1.05- 1.14	304	505	483	787	**1472**
1.15- 1.24	**685**	**989**	**1190**	**1359**	
1.25- 1.34	99	784		**1289**	
1.35- 1.44	458	557	**1242**	**1088**	**1546**
1.45- 1.54	**466**	**924**			**201**
1.55- 1.64	72	538	1023	**1708**	**2**
1.65- 1.74	202	274	740	1095	1297
1.75- 1.84	**329**	531	996	1198	
1.85- 1.94	58	**387**	603	1069	
1.95- 2.04	**273**	331	589	661	
2.05- 2.14	45	318	660	862	934
2.15- 2.24	**265**	310	376	705	970
2.25- 2.34	33	298	583	641	
2.35- 2.44	101	134	343	616	
2.45- 2.54	**196**	297	**399**	444	
2.55- 2.64	13	**209**	330	595	640
2.65- 2.74	**163**	176	310	343	506
2.75- 2.84	2	165	372	473	
2.85- 2.94	15	17	178	374	
2.95- 3.04	**129**	144	180	193	
3.05- 3.14	5	**134**	146	309	322
3.15- 3.24	**47**	52	149	151	198
3.25- 3.34	12	59	181	196	
3.35- 3.44	0	12	64	193	
3.45- 3.54	**221**	**221**	59	64	
3.55- 3.64	5	**226**	**233**	**280**	285
3.65- 3.74	**16**	21	**226**	**238**	**254**
3.75- 3.84	11	27	242	**253**	
3.85- 3.94	0	11	32	32	
3.95- 4.04	**82**	**82**	27	109	114
4.05- 4.14	0	**82**	93	93	96
4.15- 4.24	**3**	3	**82**	85	
4.25- 4.34	0	3	85	**85**	
4.35- 4.44	0	0	3	3	6
4.45- 4.54	**3**	3	3	4	4
4.55- 4.64	1	**4**	4	4	6
4.65- 4.74	0	1	4	4	
4.75- 4.84	0	0	1	1	
4.85- 4.94	0	0	0	8	9
4.95- 5.04	**8**	**8**	**8**	8	8
5.05- 5.14	0	8	8	8	
5.15- 5.24	0	0	8	9	
5.25- 5.34	1	1	1	1	

*Distribuție reprodusă după A. Bowley, **Elements of Statistics**, p. 97.*

Scanned with OKEN Scanner



Tabelul 50

Graficul 29. Distribuție de salarii în raport cu diferite scări de intervale.

56. *Antimod*. În legătură cu distribuțiile în formă de U, prof. PEARSON a introdus noțiunea de *antimod*, a cărui valoare corespunde frecvenței minime. Calculul antimodului se face fie luând jumătatea intervalului antimodal, fie calculându-l în mod mai precis cu ajutorul formulei (11). Formula (10) trebuie însă modificată, dacă vrem să o întrebuințăm la calcularea antimodului:

$$A = x_i - \frac{n_{i+1} - n_{i-1}}{n_{i+1} + n_{i-1}} \cdot \frac{d}{2}. \qquad (13)$$

Astfel, valoarea antimodului corespunzător distribuției din tabelul 36 va fi:

$$A = 5 + \frac{25}{49} \cdot \frac{1}{2} = 5{,}26$$

dacă facem uz de formula (10), și

$$A = 5 + \frac{25}{67} \cdot \frac{1}{2} = 5{,}19$$

după formula (13).

57. *Comparaţie între cele trei valori tipice fundamentale.* Valorile tipice descrise mai sus sunt cele care se întrebuinţează în mod permanent pentru a reprezenta o distribuţie de frecvenţă printr-o singură valoare caracteristică. Trebuie totuşi să accentuăm avantajele şi dezavantajele pe care le prezintă fiecare dintre acestea, pentru ca să nu fim înşelaţi în întrebuinţarea lor.

Observăm în primul rând că, dacă distribuţia este simetrică, toate valorile tipice studiate trebuie în mod teoretic să se confunde şi că, prin urmare, în practică, ele vor diferi prea puţin una de alta. Media aritmetică este suficientă pentru reprezentarea unei astfel de serii. De altfel, tot ea este valoarea tipică de recomandat în orice caz. Din nefericire, calculul ei nu se poate face întotdeauna. Este de ajuns ca o distribuţie să fie deschisă la unul din capete, pentru ca media aritmetică corespunzătoare să nu mai poată fi calculată în mod satisfăcător din cauza impreciziei ce o prezintă intervalul deschis. Iată deci cum, o dată mai mult, se constată neajunsurile distribuţiilor deschise (§ 19) şi de ce în acest caz este de dorit să avem informaţii suplimentare asupra intervalelor marginale care să ne permită să atenuăm eroarea ce eventual s-ar comite asupra mediei aritmetice prin fixarea arbitrară a limitelor acestor intervale.

Mediana este valoarea tipică al cărei calcul nu este împiedicat de faptul că seria este o distribuţie deschisă şi, din această cauză, această valoare tipică este cea mai indicată pentru reprezentarea distribuţiilor din această categorie.

Dar nici mediana şi nici media aritmetică nu corespund impresiei pe care o are cel care prelucrează datele, deoarece aceste două valori nu corespund decât foarte rar elementului celui mai frecvent şi care determină deci impresia generală ce ne-o face ansamblul de date. Modul vine să înlăture acest neajuns. În cazul distribuţiilor asimetrice, calculul modului, în afară de cel al mediei aritmetice şi al medianei, este o necesitate.

Se mai impută medianei şi modului, atât timp cât sunt considerate ca valori tipice, că nu sunt prea indicate pentru acest scop, deoarece nu *depind* de toate elementele ce alcătuiesc seria respectivă sau, cu alte cuvinte, nu sunt o expresie a tipicului extras din întregul ansamblu al manifestărilor unui fenomen. Obiecţia aceasta este foarte justă, însă de obicei nici mediana, nici modul nu sunt întrebuinţate ca valori tipice, ci ele sunt calculate pentru ca, împreună cu media aritmetică, să avem o mai completă reprezentare a realităţii.

Acest neajuns al medianei şi al modului are, cu toate acestea, partea lui utilă: din cauză că aceste valori nu sunt influenţate de toate elementele seriei, ele vor fi mai indicate să reprezinte o serie în care avem motive să credem că s-au strecurat elemente excepţionale în număr neobişnuit.

Mediana şi modul sunt valori tipice speciale studiului distribuţiilor heterograde. Valoarea tipică a unei serii de valori individuale, cum este o serie în timp sau o distribuţie în spaţiu, nu poate fi în mod normal decât media aritmetică sau altele câteva pe care le vom întâlni în paragrafele următoare.

În sfârşit, pentru a evita o extindere greşită a proprietăţilor mediei aritmetice, este necesar să subliniem faptul că nici mediana, nici modul n-au - în cazul general - teoreme de adiţie (§ 46). Cu alte cuvinte, nici modul, nici

Scanned with OKEN Scanner

mediana unei distribuţii formate din contopirea altora nu pot fi calculate numai cu ajutorul medianelor sau modurilor primitive; calculul, în acest caz, trebuie făcut efectiv, pornind de la noua distribuţie.

58. *Valore poziore*. O altă valoare tipică, introdusă de şcoala italiană, este aşa-numita *valore poziore*. Termenul n-a fost încă tradus în literaturile străine, propunem totuşi pentru vocabularul român pe cel de *valoare potenţială*[1]. Această valoare, care este foarte rar întrebuinţată, se obţine în felul următor:

Se fac toate produsele dintre valori şi frecvenţele respective; valoarea căreia îi corespunde produsul maxim este valoarea potenţială a seriei (tabelul 51).

TABELUL 51

Nota obţinută la intrarea în Facultatea de Ştiinţe, de către un grup de elevi provenind din învăţământul clasic

Nota	Frecvenţe	Produsul (1) × (2)
(1)	(2)	(3)
6	243	1.458
7	253	1.771
8	233	1.864
9	210	**1.890**
10	137	1.370

Extras din A. Niceforo, ***La méthode statistique et ses applications aux sciences naturelles, aux sciences sociales et à l'art****, Paris, Giard, 1925, p.162.*

Valore poziore = 9.

Valoarea potenţială corespunde, am putea spune, valorii care are *potenţialul maxim.* În cazul examinat mai sus, nu grupa cea mai numeroasă (elevii care au obţinut nota 7) are şi potenţialul cel mai mare, ceea ce înseamnă că în general modul nu este egal cu valoarea potenţială.

Această valoare tipică îşi găseşte aplicaţii mai ales în studiul statistic al fenomenelor psihologice. În repartiţiile de averi, venituri, impozite, valoarea potenţială reprezintă clasa care, deşi poate să nu fie cea mai numeroasă, este cea mai bogată sau plăteşte cea mai mare parte din impozitul total.

Cu toată importanţa pe care o reprezintă ideea care a condus la definirea valorii potenţiale, această valoare tipică este încă prea puţin răspândită.

59. *Media geometrică*. Grupul de valori tipice întrebuinţate în statistică este mult mai variat decât reiese numai din consideraţiile de până acum. Vom cita, în primul rând, media geometrică şi media armonică pe care le vom

[1] *În urma sugestiilor dlui prof. O. Onicescu.*

Scanned with OKEN Scanner

examina succesiv. Este necesar însă să accentuăm dintr-un început faptul că ambele aceste noţiuni au un caracter matematic pronunţat. Însăşi înglobarea lor în categoria valorilor tipice nu mi se pare destul de justificată (poate cu o uşoară excepţie pentru media geometrică). Este greu de redat în mod intuitiv calitatea tipică a seriei reprezentată de aceste valori, deşi calculul lor se sprijină pe ansamblul elementelor constitutive. Lucrurile nu trebuie înţelese greşit; dacă aceste valori nu sunt prea utile ca valori tipice, aduc însă reale servicii metodei statistice din alte puncte de vedere. Dacă vorbim despre ele acum este mai mult pentru a respecta o tradiţie în această direcţie.

Media geometrică a unui ansamblu de valori $x_1, x_2, \ldots, x_i$ este dată de formula:

$$G = \sqrt[i]{x_1 x_2 x_3 \ldots x_i},$$

de unde

$$\log G = \frac{\sum \log x}{i} \qquad (14)$$

sau în cazul unei distribuţii heterograde:

$$\log G = \frac{\sum n \log x}{N}. \qquad (14 \text{ bis})$$

Relaţiile acestea ne permit calculul efectiv al mediei geometrice, bineînţeles, servindu-ne de o tablă de logaritmi.

Întrebuinţarea mediei geometrice este legată mai mult de seriile în timp ale fenomenelor ce cresc sau descresc în mod *progresiv*.

Deşi vom analiza pe larg, mai târziu, seriile în timp (Cap. *Analiza seriilor în timp),* când vom vorbi şi despre acest tip de fenomene, este suficient să spunem pentru moment că fiecare valoare în seria statistică corespunzătoare unui atare fenomen reprezintă, faţă de valoarea precedentă, o majorare sau o diminuare relativă (la sută) aproape constantă de-a lungul seriei.

Să presupunem că avem de a face cu un fenomen a cărui intensitate creşte în cursul unui anumit interval de timp cu 20%; dacă valoarea iniţială este egală cu 100, de pildă, valorile următoare vor fi:

$$\begin{aligned} 100 &\times 1{,}2 = 120 \\ 120 &\times 1{,}2 = 144 \\ 144 &\times 1{,}2 = 172{,}8 \\ 172{,}8 &\times 1{,}2 = 207{,}4 \\ 207{,}4 &\times 1{,}2 = 248{,}8 \\ &\ldots\ldots\ldots\ldots\ldots \end{aligned}$$

Scanned with OKEN Scanner

În general, pornind de la valoarea P_0, vom obține șirul:

$$P_0, P_0r, P_0r^2, \ldots\ldots, P_0r^n$$

cu alte cuvinte, o progresie geometrică.

Astfel de fenomene se întâlnesc destul de des în mai toate domeniile de activitate și oglindesc în ele dezvoltarea din ce în ce mai pronunțată a populațiilor, a producției industriale, a intensității schimbului comercial etc. Să considerăm, de exemplu, creșterea populației din Statele Unite ale Americii (tabelul 52). În acest tabel, cifrele din coloana (3) reprezintă creșterea absolută a populației de la un recensământ la altul. Cele din coloana (4) reprezintă aceeași creștere raportată la cifra recensământului precedent. Astfel, creșterea de la 1840 la 1850 a fost de 6 123 000, care, raportată la cifra populației din 1840, ne dă o creștere relativă de 35,9%.

TABELUL 52

Populația Statelor Unite ale Americii după recensămintele succesive

Anii	Populația (în mii)	Creșterea	
		absolută	relativă
1790	3 929	-	-
1800	5 308	1 379	35,1
1810	7 240	1 932	36,4
1820	9 638	2 398	33,1
1830	12 866	3 228	33,5
1840	17 069	4 203	32,7
1850	23 192	6 123	35,9
1860	31 443	8 251	35,6
1870	38 558	7 115	22,6
1880	50 156	11 598	30,1
1890	62 948	12 792	25,5
1900	75 995	13 047	20,7
1910	91 972	15 977	21,0
1920	105 711	13 739	14,9
1930	122 775	17 064	16,1

Extras din ***Fifteenth Census of the U. S.****, Population, 1930, vol. I, p. 6.*

Se vede cum creșterea relativă, deși variază și ea, variația ei se face în mod mai lent, iar pentru o bună perioadă de timp ea poate fi privită ca o constantă statistică. Singur recensământul din 1920 prezintă o creștere excepțional de mică, dar care își găsește explicația în faptul că acest

Scanned with OKEN Scanner

recensământ a urmat războiului mondial care, ca toate războaiele, a constituit un element de distrugere a populației și de împiedicare a dezvoltării ei normale.

60. *Aplicațiile mediei geometrice*. Dacă se consideră un șir de valori consecutive ale unei serii de tipul descris:

$$P_0r^{j-2},\ P_0r^{j-1},\ P_0r^{j},\ P_0r^{j+1},\ P_0r^{j+2}$$

valoarea centrală este media geometrică a șirului:

$$\sqrt[5]{P_0r^{j-2},P_0r^{j-1},P_0r^{j},P_0r^{j+1},P_0r^{j+2}} = P_0r^{j}.$$

De aici, două constatări:

1. Valoarea care reprezintă intensitatea manifestărilor unui fenomen progresiv, în cursul unei perioade oarecare de timp, trebuind să fie întrucâtva o *valoare mijlocie*, vom adopta pentru aceasta media geometrică a șirului.
2. Atunci când nu cunoaștem un termen intermediar al unei astfel de serii, media geometrică a termenilor care-l încadrează în timp este o valoare mai judicioasă de atribuit termenului ignorat decât media aritmetică a aceloraşi termeni.

Să exemplificăm aceste considerații și, în acest scop, să considerăm producția mondială de huilă pentru anii 1891-1910 (tabelul 53).

Astfel, producția mijlocie pentru perioada 1891-1910 va fi mai logic reprezentată de media geometrică a seriei respective:

$$G = \text{antilog}\ \frac{56{,}9295}{20} = \text{antilog}\ 2{,}84647 = 702{,}2$$

decât de media aritmetică:

$$M = \frac{14\,534}{20} = 726{,}7\,.$$

Și dacă în exemplul de față diferența nu este chiar așa de accentuată este din cauză că proporția în care a crescut de la un an la altul producția mondială de huilă este foarte mică.

Să trecem acum la cea de-a doua observație făcută. Să presupunem, de pildă, că am fi ignorat cifra populației din Statele Unite pentru anul 1850. Dacă atribuim atunci termenului corespunzător valoarea mediei aritmetice a populației din 1840 și 1860, vom avea:

$$M = \frac{17\,069 + 31443}{2} = 24\,256 \text{ mii}$$

pe când media geometrică a aceloraşi termeni va fi :

$$G = \sqrt{17\,069 \times 31443} = 23\,167 \text{ mii}$$

Scanned with OKEN Scanner

TABELUL 53

Producţia mondială de huilă în milioane de tone metrice

(Evaluări)

Anii	Producţia	Logaritmi	Totaluri parţiale	
			Producţie	Logaritmi
1891	494	2,6937		
1892	498	2,6972		
1893	487	2,6875		
1894	509	2,7067		
1895	536	2,7292	2 524	13,5143
1896	551	2,7412		
1897	577	2,7612		
1898	608	2,7839		
1899	666	2,8235		
1900	700	2,8451	3 102	13,9549
1901	718	2,8561		
1902	735	2,8663		
1903	807	2,9069		
1904	811	2,9090		
1905	858	2,9335	3 929	14,4718
1906	924	2,9657		
1907	1 020	3,0086		
1908	966	2,9850		
1909	1 010	3,0043		
1910	1 059	3,0249	4 979	14,9885
	14 534	**56,9295**		

Extras din P. Gemäling, ***Statistiques choisies et annotées****, Paris, Sirey, 1926, p.53.*

Media geometrică ne permite să ne apropiem în acest caz de valoarea adevărată, 23.192 mii, într-un mod cum nu se poate mai satisfăcător.

O altă problemă la care întrebuinţarea mediei geometrice constituie un procedeu util nu este decât o faţă mai complicată a chestiunii precedente.

Recensămintele populaţiilor se fac de obicei din 10 în 10 ani. Între timp, cifra populaţiei ne este necunoscută. Nu putem avea decât evaluări ale ei, făcute cu ajutorul datelor relative la mişcarea populaţiei. Se obţin totuşi valori presupuse pentru numărul populaţiei an cu an, în chipul următor:

Dacă P_0 este cifra populaţiei din anul 1840, iar r, creşterea proporţională de la un an la altul, populaţiile *probabile* din 1841, 1842, 1843, ..., 1850 vor fi egale cu P_0r, P_0r^2, P_0r^3,..., P_0r^{10}.

Cum în exemplul de mai sus $P_0 = 17\ 069$ şi $P_0r^{10} = 23\ 192$, urmează că:

Scanned with OKEN Scanner

$$r^{10} = \frac{23\,192}{17\,069} = 1{,}359$$

şi deci:

$$\log r = \frac{\log 1{,}359}{10} = 0{,}013322$$

de unde:

$$r = 1{,}03114.$$

Aşadar, populaţia Statelor Unite a crescut anual, în perioada 1840-1850, cu 3,114%.

Dacă am adopta pentru această evaluare media aritmetică a creşterii totale:

$$r' = \frac{35{,}9}{10} = 3{,}59\%$$

am obţine o valoare cu totul diferită de precedenta şi, bineînţeles, eronată.

TABELUL 54

Populaţia probabilă a Statelor Unite în perioada 1840-1850[1]

Anii	Calcul bazat pe media geometrică	Calcul bazat pe media aritmetică
1840	17 069	17 069
1841	17 601	17 682
1842	18 149	18 317
1843	18 714	18 974
1844	19 207	19 655
1845	19 897	20 361
1846	20 517	21 092
1847	21 156	21 849
1848	21 815	22 633
1849	22 494	23 446
1850	23 194	24 288

Populaţia probabilă, calculată după ambele rezultate, este cea din tabelul 54. Eroarea semnalată este evidentă pe acest tabel; calculul prin media aritmetică ne dă pentru anul 1850 un surplus de populaţie egal cu un milion.

[1] *Pentru un calcul mai precis şi, totodată, mai comod, a se vedea capitolul* Analiza seriilor în timp. *Multe din chestiunile atinse în paragrafele de mai sus ne vor apărea mai clare în urma expunerii din acel capitol.*

Scanned with OKEN Scanner

61. *Proprietăţile mediei geometrice.* Media geometrică este întotdeauna mai mică sau cel mult egală cu media aritmetică a aceluiaşi ansamblu de valori:

$$G \leq M .$$

Diferenţa dintre aceste două valori este cu atât mai mică cu cât valorile ansamblului respectiv diferă mai puţin între ele; egalitatea nu se produce decât dacă toate valorile ansamblului sunt egale. Prin urmare, întrebuinţarea mediei geometrice ca valoare tipică ne conduce la o subestimare a ordinului de mărime.

Ca şi media aritmetică, media geometrică admite o teoremă de adiţie. Media geometrică a unei serii formată din contopirea mai multor grupuri, ale căror medii geometrice sunt G_1, G_2, ..., G_i, fiecare conţinând respectiv n_1, n_2, ..., n_i elemente, este dată de formula:

$$\log G = \frac{\sum n \log G}{\sum n} \tag{15}$$

care devine:

$$\log G = \frac{\sum \log G}{i} \tag{16}$$

în cazul când grupurile componente sunt formate din acelaşi număr de elemente.

Astfel, din tabelul 53 deducem mediile geometrice ale grupelor succesive de cinci ani:

G_1 = antilog 2,70286 = 504,5
G_2 = antilog 2,79098 = 617,9
G_3 = antilog 2,89436 = 784,1
G_4 = antilog 2,99770 = 994,7

Deci:

$$\log G = \frac{2{,}70286 + 2{,}79098 + 2{,}89436 + 2{,}99770}{4} = \frac{11{,}38590}{4} = 2{,}84647$$

rezultat care concordă cu cel obţinut pe cale directă.

Trebuie remarcat că acest calcul nu a presupus decât cunoaşterea mediilor G_1, G_2, ... şi a numărului de elemente ce compun grupurile, iar nu şi a valorilor individuale.

62. *Media armonică.* Media armonică este şi mai rar întrebuinţată decât media geometrică. Definiţia ei este dată de formula următoare:

Scanned with OKEN Scanner

$$H = \frac{N}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \ldots + \frac{n_i}{x_i}} \qquad (17)$$

unde notaţiile sunt cele întrebuinţate mai sus.

TABELUL 55

Distribuţia unui grup de flori de *Caltha palustris* în raport cu numărul petalelor

Numărul petalelor x	Frecvenţe n	Câtul $\frac{n}{x}$
(1)	(2)	(3)
5	300	60
6	87	14,5
7	24	3,4
8	4	0,5
	415	**78,4**

*Extras din E. Coquidé, **op. cit.**, p. 103.*

Calculul acestei valori este destul de anevoios, inversele numerelor considerate necesitând fiecare câte o operaţie de împărţire. Există şi table pentru aflarea directă a inverselor numerelor naturale, însă acestea nu cuprind toate cazurile ce se pot prezenta. Tabelul 55 constituie un exemplu de calcul al mediei armonice.

După formula (17), media armonică a acestei distribuţii va fi:

$$H = \frac{415}{78,4} = 5,29.$$

Media aritmetică şi media geometrică ale aceleiaşi distribuţii vor fi:

$$M = 5,35 \qquad G = 5,32.$$

Ordinea de mărime întâlnită aici este generală, adică vom avea întotdeauna:

$$H \leq G \leq M \qquad (18)$$

egalitatea neexistând decât când toate valorile ce compun seria sunt egale.

Scanned with OKEN Scanner

Din cauza acestei inegalități, este poate mai bine să adoptăm ca valoare tipică a unei distribuții în formă de *J* media armonică a acesteia, decât vreuna din celelalte două medii. În acest caz, valoarea mediei armonice este mai apropiată și de modul și de mediana corespunzătoare.

Domeniul în care media armonică își găsește însă cea mai bună aplicație este calculul indicilor bursieri. Să presupunem că toate valorile cu venit fix au fost reduse la valoarea nominală de 100 și la același procent nominal, 5%, de exemplu. Astfel calculate, cotele a trei obligațiuni sunt:

$$152, \quad 114, \quad 78$$

Plasamentele de capital în aceste trei valori aduc, respectiv, următoarele dobânzi efective:

$$\frac{500}{152} = 3{,}29\% \quad \frac{500}{114} = 4{,}39\% \quad \frac{500}{78} = 6{,}41\%.$$

Un indice reprezentativ al valorilor acestora poate fi calculat fie pornind de la cursurile lor în bursă, fie de la rentabilitățile lor[1]. Cum în materie de plasament interesează în primul rând rentabilitatea efectivă a unei valori bursabile, se poate adopta pentru indicele căutat media aritmetică a procentelor de mai sus, care este 4,70%.

Acest mod de a reprezenta un ansamblu de valori cu venit fix, deși logic, nu este de recomandat, deoarece bursele nu ne dau procentul efectiv al unei valori, ci cota ei. Este foarte natural să căutăm să reprezentăm atunci ansamblul de valori considerat printr-o valoare care să ne raporteze procentul mediu calculat, 4,70%. Or, această valoare este tocmai media armonică a cursurilor de mai sus:

$$H = \frac{3}{\frac{1}{152} + \frac{1}{114} + \frac{1}{78}} = 106{,}5\,.$$

Iată deci cum putem avea un indice bursier rațional, fără ca să calculăm în prealabil procentul efectiv.

Media aritmetică a acelorași valori - 114,67 - ar corespunde unui procent mai mic (4,36%).

63. *Medii de ordin superior. Momente.* Din considerațiile precedente se desprinde utilitatea noțiunii de valoare tipică sau mijlocie, precum și diferitele procedee prin care putem obține cele mai uzitate valori ce corespund acestei noțiuni.

S-au mai imaginat și alte valori medii (o descriere completă a lor ar putea face singură obiectul unui tratat) care sunt folosite în diferitele domenii în care

[1] *Vezi Cap.* Numere indice.

Scanned with OKEN Scanner

instrumentul matematic poate fi util. Chiar metoda statistică face uz şi de alte valori medii. Caracterele distribuţiilor pe care aceste valori le oglindesc sunt însă cu totul altele decât cele întâlnite până acum. Le vom examina la timp. Din motive de analogie a procedeului de calcul, ne vom ocupa însă aici de cele mai importante.

Acestea sunt mediile de ordin superior.

Astfel, *media pătratică* este definită de relaţia:

$$M_2 = \sqrt{\frac{\sum nx^2}{\sum n}} \tag{19}$$

media cubică, de:

$$M_3 = \sqrt[3]{\frac{\sum nx^3}{\sum n}} \tag{20}$$

şi, în general, *media de ordinul p* este dată de formula:

$$M_p = \sqrt[p]{\frac{\sum nx^p}{\sum n}}. \tag{21}$$

Se vede că, pentru aflarea acestor medii, trebuie să calculăm în prealabil expresii de forma:

$$m_p = \frac{\sum nx^p}{\sum n} \tag{22}$$

care nu sunt altceva decât mediile aritmetice ale diferitelor puteri ale valorilor ce constituie ansamblul respectiv.

Expresiile (22) poartă numele de *momentele distribuţiei*. (Momentul de ordinul 1 este chiar media aritmetică.) Importanţa lor pentru studiul distribuţiilor heterograde este capitală. Ansamblul momentelor caracterizează distribuţia respectivă. Din această cauză, mai toate cercetările moderne de statistică matematică tind să descopere modul în care aceste momente se comportă în practică, adică legea după care sunt influenţate de hazard.

Tabelul 56 ne arată care este procedeul de calcul al momentelor şi deci al mediilor de diferite ordine ale unei distribuţii. Procedeul acesta este cel direct şi nu este întrebuinţat în practică decât atunci când seria este formată dintr-un număr prea restrâns de elemente. Vom vedea în capitolul următor care este procedeul de calcul simplificat.

Scanned with OKEN Scanner

TABELUL 56

Calculul mediilor de ordin superior

Numărul razelor x	Frecvenţe n	Pătrate x^2	Cuburi x^3	Produsele	
				nx^2	nx^3
5	1	25	125	25	125
6	5	36	216	180	1 080
7	9	49	343	441	3 087
8	22	64	512	1 408	11 264
9	38	81	729	3 078	27 708
10	62	100	1 000	6 200	62 000
11	61	121	1 331	7 381	81 191
12	29	144	1 728	4 176	50 112
13	14	169	2 197	2 366	30 758
14	4	196	2 744	784	10 976
15	4	225	3 375	900	13 500
	249			**26 939**	**291 795**

Tabelul 27

$$m_2 = \frac{26\,939}{249} = 108{,}1887 \qquad M_2 = \sqrt{m_2} = 10{,}4$$

$$m_3 = \frac{291795}{249} = 1171{,}867 \qquad M_3 = \sqrt[3]{m_3} = 10{,}6$$

64. *Proprietăţile esenţiale ale valorilor tipice*. Expresiile de valori medii pot fi multiplicate la infinit, nu însă fără a fi supuse unor condiţii generale. Aceste condiţii, puse în evidenţă de matematicianul CAUCHY, sunt următoarele:

1. Orice valoare medie trebuie să fie cuprinsă, ca mărime, între cea mai mică şi cea mai mare valoare a ansamblului respectiv (ea va fi deci egală cu valoarea comună a elementelor constitutive, în cazul când acestea toate vor fi egale între ele).
2. Valoarea medie trebuie să rămână neschimbată dacă frecvenţele seriei se înmulţesc sau se împart cu acelaşi factor.

Ambele aceste condiţii sunt îndeplinite de toate valorile tipice considerate în acest capitol. Şi este bine să accentuăm concluzia care se desprinde din cea de a doua condiţie: *valorile tipice ale unei distribuţii sunt aceleaşi fie că sunt calculate pornind de la frecvenţele absolute, fie de la frecvenţele relative.*

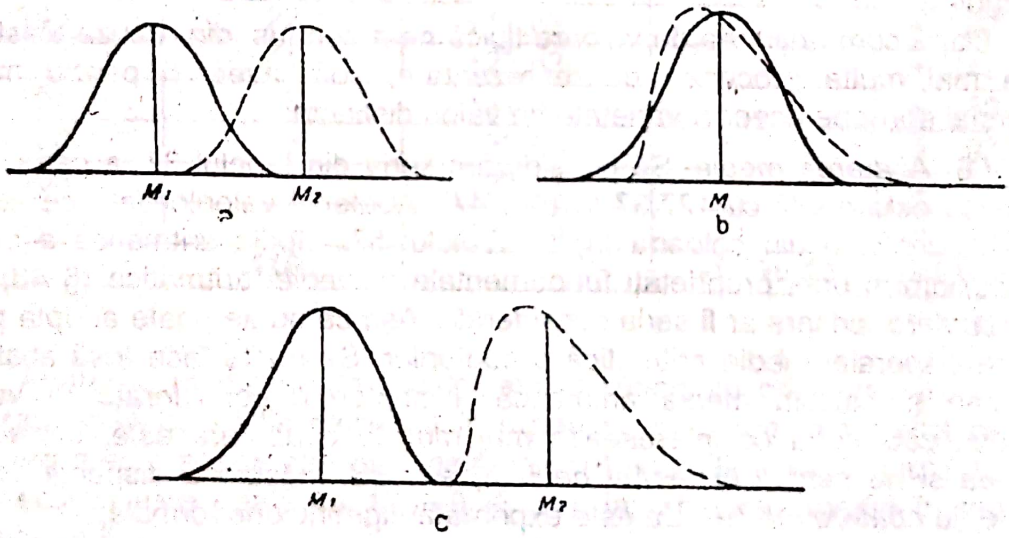
Scanned with OKEN Scanner

Capitolul V

Dispersie, asimetrie

65. *Importanța statistică a noțiunii de dispersie.* Oricare din valorile tipice unei distribuții nu ne oferă decât concretizarea printr-un număr a unui singur aspect al acesteia. Acest aspect este, după cum am văzut, ordinul de mărime mijlociu al elementelor componente. Orice distribuție are însă și o organizare interioară care constă din modul cum valorile respective se repartizează în jurul valorii tipice și care poate diferi de la un fenomen la altul.

Am arătat la începutul acestei lucrări care este importanța unui studiu îndreptat în această direcție. Să analizăm acum mai de aproape lucrurile.



Graficul 30. Cazurile de diferențiere a distribuțiilor heterograde.

O valoare tipică, media aritmetică, de pildă, nu poate fi un criteriu unic de diferențiere a două distribuții. Într-adevăr, lăsând deocamdată deoparte forma distribuției și considerând două distribuții de același tip, acestea se pot deosebi, după cum se vede în graficul 30:

1. Numai prin valoarea medie (a);
2. Numai prin organizarea interioară (b);

Scanned with OKEN Scanner

3. Şi prin valoarea medie, şi prin organizarea interioară (c).

Această organizare interioară depinde în primul rând de mărimea diferenţei între elementele ce au cea mai mare şi cea mai mică valoare. Un număr care să ne măsoare numai *întinderea* câmpului corespunzător unei distribuţii este chiar această diferenţă care poartă numele de *oscilaţie*. Astfel, oscilaţia numărului de morţi la mia de locuitori va fi, pentru România (tabelul 45), egală cu 28,1 - 22,9 = 5,2. Pentru înălţimea barometrică la Southampton, oscilaţia va fi egală cu 31,05 - 28,45 = 2,6 degete (tabelul 32).

Dar organizarea interioară a unei distribuţii nu este cuprinsă numai în ideea de oscilaţie; modul cum se repartizează elementele împrejurul valorii tipice este o faţă importantă a acesteia şi noţiunea la care dă naştere poartă numele de *dispersie*.

Dispersia este una din caracteristicile fundamentale ale manifestărilor fenomenelor colective. Ea corespunde gradului de variabilitate a speciilor biologice, de stabilitate a fenomenelor sociale, iar atunci când se referă la o serie de măsurători ale aceleiaşi mărimi, măsoară precizia cu care s-au efectuat acestea. Pentru o interpretare numerică a dispersiei, oscilaţia nu este de ajuns. Trebuie căutată o valoare care să aibă caracterul unei valori medii, raportată de data aceasta la abaterile elementelor faţă de valoarea tipică a seriei, căci, în definitiv, ansamblul acestor abateri defineşte prin structura lui dispersia seriei, iar noi căutăm o valoare care să ne reprezinte ceea ce este tipic în acest ansamblu.

După cum noţiunea de valoare tipică ne-a condus, din cauza elasticităţii ei, la mai multe procedee de reprezentare, vom avea şi pentru măsura dispersiei aproape aceeaşi varietate de valori distincte.

66. *Abaterea medie*. Să considerăm seria din tabelul 31, a cărui medie aritmetică este egală cu 127,37 livre (§ 47). Abaterile valorilor faţă de această valoare sunt cele din coloana (3) a tabelului 57. Media aritmetică a acestor valori, conform unei proprietăţi fundamentale a mediei aritmetice (§ 48), este egală cu zero, oricare ar fi seria considerată. Aşa că nu se poate adopta pentru măsura dispersiei media aritmetică a abaterilor. Se poate face însă abstracţie de semn şi calcula media aritmetică a abaterilor considerate în valoare absolută, căci ceea ce interesează mai mult la o abatere este, într-adevăr, mărimea şi nu sensul ei. Astfel definită, această măsură a dispersiei poartă numele de *abatere medie*[1]. Ea este exprimată algebric prin formula:

$$e_M = \frac{\sum n\,|\xi|}{N}, \tag{23}$$

unde $|\xi|$ înseamnă valoarea fără semn a abaterii ξ.

În cazul distribuţiei considerate, avem:

[1] Ecart moyen, *în franceză*; mean deviation, *în terminologia engleză*.

$$e_M = \frac{1826{,}14 + 1825{,}22}{416} = 8{,}78\ ^{1}.$$

TABELUL 57

Calculul abaterii medii a distribuției unui grup de adulți în raport cu greutatea

Mijlocul intervalului x	Frecvențe n	Abaterile față de medie ξ	Produsul $n\xi$	
(1)	(2)	(3)	(4)	
100	3	- 27,37	- 82,11	
105	6	- 22,37	- 134,22	
110	23	- 17,37	- 399,51	
115	48	- 12,37	- 593,76	
120	56	- 7,37	- 412,72	
125	86	- 2,37	- 203,82	
130	78	+ 2,63		+ 205,14
135	45	+ 7,63		+ 343,35
140	30	+12,63		+ 378,90
145	22	+17,63		+ 387,86
150	11	+22,63		+ 248,93
155	3	+27,63		+ 82,89
160	3	+32,63		+ 97,89
165	1	+37,63		+ 37,63
170	1	+42,63		+ 42,63
	416		**-1 826,14**	**+1 825,22**

Tabelul 31

Abaterea medie poate fi calculată în raport cu orice valoare tipică. E preferabil însă ca să fie calculată în raport cu *mediana, din cauză că astfel abaterea medie atinge cea mai mică valoare posibilă a sa*. În cazul tabelului 57, mediana este egală cu 126,69 (§ 51), iar abaterea medie în raport cu aceasta va fi:

$$e_{Me} = 8{,}73.$$

[1] *Dacă în cazul tabelului 57 media aritmetică a abaterilor în raport cu media aritmetică a distribuției nu este egală chiar cu 0 este din cauză că aceasta din urmă n-a fost calculată decât cu aproximație de 0,01. Media aritmetică a abaterilor în valoare relativă este în acest caz 0,0022, valoare care trebuie privită ca egală cu 0 în ordinul de aproximație adoptat.*

Scanned with OKEN Scanner

Diferența între abaterile medii în raport cu media aritmetică și în raport cu mediana este, evident, cu atât mai accentuată cu cât distribuția este mai asimetrică.

Procedeul de calcul al abaterii medii este destul de anevoios; el poate fi în parte simplificat însă cu ajutorul unor complicații care întrec simplificarea ce se obține.

67. *Abatere-tip. Fluctuație.* Un alt mod de a măsura dispersia unei distribuții este dat de *abaterea-tip*[1]. Definiția abaterii tip este concretizată în formula:

$$\sigma = \sqrt{\mu_2} = \sqrt{\frac{\sum n\,\xi^2}{N}}, \tag{24}$$

adică este egală cu media pătratică a abaterilor în raport cu media aritmetică a seriei. În formula de mai sus μ_2 este momentul de ordinul al doilea în raport cu media aritmetică și este numit de unii autori cu termenul de *fluctuație.*

În comparație cu abaterea medie, valoarea abaterii tip este influențată de o ponderare ascunsă, consecința faptului că fiecare abatere nu intervine direct în calcul, ci prin pătratul ei. Din această cauză, abaterile mari - în special cele excepționale - au o contribuție mai mare decât celelalte abateri la formarea acestei valori. Și, ca o consecință imediată a acestui fapt, deducem că valoarea abaterii tip va fi întotdeauna mai mare decât valoarea abaterii medii.

Cu toate acestea, întrebuințarea acestei valori este relativ cea mai răspândită, fapt care este în mare parte justificat și de importanța teoretică a lui σ.

Pentru calculul abaterii-tip, în afară de procedeul direct - care, după atâtea exemple, nu mai are nevoie să fie lămurit și care nu este decât foarte rar întrebuințat - mai există un procedeu simplificat. Acesta din urmă prezintă o perfectă analogie cu procedeul simplificat pentru calculul mediei aritmetice, cu altele pe care le vom întâlni în legătură cu momentele de ordin superior și este bazat pe o proprietate generală a momentelor. Metoda de calcul constă din schimbarea originii și din aplicarea relației:

$$\mu_2 = m_2(a) - \overline{m_1(a)}^2, \tag{25}$$

unde $m_2(a)$, $m_1(a)$ sunt momentele de ordinul 2 și 1, în raport cu valoarea a adoptată ca origine[2]. Calculul amănunțit se vede în tabelul 58.

---

[1] Ecart quadratique moyen *sau* écart type, *în franceză;* standard deviation, *în engleză.*

[2] *$m_1(a)$ nu este altceva decât M(a) din formula (6 bis).*

TABELUL 58

Calculul fluctuației pentru distribuția unui grup de boabe de fasole, în raport cu lungimea

Lungimea în mm	Mijlocul intervalului x	Frecvențe n	Abateri în raport cu 12 α	Abateri la pătrat α^2	Produsul $n\alpha$	Produsul $n\alpha^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
7,5- 8,5	8	1	- 4	16	- 4	16
8,5- 9,5	9	2	- 3	9	- 6	18
9,5-10,5	10	23	- 2	4	- 46	92
10,5-11,5	11	108	- 1	1	- 108	108
11,5-12,5	12	167	-	-	**- 164**	-
12,5-13,5	13	106	+1	1	+106	106
13,5-14,5	14	33	+2	4	+ 66	132
14,5-15,5	15	7	+3	9	+ 21	63
15,5-16,5	16	1	+4	16	+ 4	16
		448			**+197**	**551**

Datele au fost obținute de H. de Vries și sunt reproduse după G. K. Constantinescu, ***Ereditatea experimentală****, 1924, p. 24.*

Din acest tabel deducem:

$$m_1(a) = \frac{197 - 164}{448} = 0{,}074 \qquad m_2(a) = \frac{551}{448} = 1{,}2299$$

$$\mu_2 = 1{,}2299 - 0{,}0055 = 1{,}2244$$

și deci:

$$\sigma = \sqrt{\mu_2} = 1{,}11 \ .$$

Avantajul acestei metode de calcul nu constă numai din faptul că ne permite calcularea momentului de ordinul al doilea în raport cu media aritmetică fără ca să cunoaștem această din urmă valoare, ci și din acela că putem obține în același timp și valoarea momentului de ordinul întâi (media aritmetică). Într-adevăr, avem după (6 bis):

$$M = 12 + 0{,}074 = 12{,}074.$$

Scanned with OKEN Scanner

Dacă abaterile sunt măsurate în intervale - ceea ce este posibil numai în cazul când acestea sunt egale - calculul se ușurează mult. Formula (25) trebuie însă înlocuită prin următoarea:

$$\mu_2 = \left[m_2(a) - \overline{m_1(a)}^2\right] \times d^2. \qquad (26)$$

TABELUL 59

Numărul bătăilor pulsului pe minut la un grup de condamnați

Numărul bătăilor pe minut	Mijlocul intervalelor x	Frecvențe n	Abateri în raport cu 78,5 α	Abateri la pătrat α^2	Produsul $n\alpha$	Produsul $n\alpha^2$
(1)	(2)	(3)	(4)	(5)	(6)	
45 - 48	46,5	2	- 8	64	- 16	128
49 - 52	50,5	5	- 7	49	- 35	245
53 - 56	54,5	17	- 6	36	- 102	612
57 - 60	58,5	57	- 5	25	- 285	1 425
61 - 64	62,5	90	- 4	16	- 360	1 440
65 - 68	66,5	150	- 3	9	- 450	1 350
69 - 72	70,5	120	- 2	4	- 240	480
73 - 76	74,5	131	- 1	1	- 131	131
77 - 80	78,5	109	-	-	**-1 619**	-
81 - 84	82,5	86	+ 1	1	+ 86	86
85 - 88	86,5	62	+ 2	4	+ 124	248
89 - 92	90,5	42	+ 3	9	+ 126	378
93 - 96	94,5	15	+ 4	16	+ 60	240
97 - 100	98,5	18	+ 5	25	+ 90	450
101 - 104	102,5	9	+ 6	36	+ 54	324
105 - 108	106,5	5	+ 7	49	+ 35	245
109 - 112	110,5	3	+ 8	64	+ 24	192
113 - 116	114,5	3	+ 9	81	+ 27	243
	-	**924**	-	-	**+ 626**	**8 217**

Extras din M. H. Whiting, ***A Study of Criminal Anthropometry****, Biometrika, vol. XI, 1915, p. 1.*

Tabelul 59 ne arată care este mecanismul aplicării acestei formule. Din ultimele două coloane deducem:

$$m_1(a) = \frac{-1619 + 626}{924} = -1{,}075 \qquad m_2(a) = \frac{8217}{924} = 8{,}8929$$

și deci:

$$\sigma = \sqrt{m_2(a) - \overline{m_1(a)^2}} \text{ intervale}$$
$$= 2{,}78 \text{ intervale} = 2{,}78 \times 4 = 11{,}12$$
$$M = 78{,}5 + 4 \times (-1{,}075) = 74{,}2.$$

Valoarea abaterii medii a distribuției din tabelul 59 este egală cu 8,83. Se verifică astfel proprietatea de care vorbeam mai sus, adică:

$$\sigma \geq e_M .$$

68. *Proprietățile abaterii-tip.* Sunt câteva proprietăți ale abaterii-tip care trebuie neapărat subliniate pentru ca importanța acestei valori, așa de des întrebuințată în statistică, să fie deplin înțeleasă.

Din relația (25) se deduce:

$$m_2(a) = \mu_2 + \overline{m_1(a)^2}$$

ceea ce ne arată că momentul în raport cu o valoare oarecare este întotdeauna mai mare decât μ_2 - momentul în raport cu media aritmetică a distribuției. Sau, cu alte cuvinte, fluctuația constituie un minimum pentru momentul de ordinul al doilea. Abaterea-tip deci va avea cea mai mică valoare atunci când va fi calculată în raport cu media aritmetică, ceea ce de altfel se face întotdeauna.

O altă proprietate a abaterii-tip este faptul că admite o teoremă de adiție.

Dacă două serii ale căror medii aritmetice și abateri-tip sunt respectiv M_1, M_2; σ_1, σ_2 se contopesc în una singură, atunci abaterea-tip a acesteia din urmă va fi dată de formula:

$$\sigma^2 = \frac{N_1(\sigma_1^2 + d_1^2) + N_2(\sigma_2^2 + d_2^2)}{N_1 + N_2} \qquad (27)$$

unde N_1, N_2 reprezintă numărul elementelor cuprinse în seriile componente, iar:

$$d_1 = M - M_1 \qquad d_2 = M - M_2$$

M fiind media aritmetică a seriei rezultante:

$$M = \frac{N_1 M_1 + N_2 M_2}{N_1 + N_2}.$$

Scanned with OKEN Scanner

Proprietatea exprimată prin relaţia (27) este generală, adică se poate extinde la cazul contopirii mai multor serii într-una singură, însă aceasta având mai mult o importanţă teoretică, nu insistăm mai mult asupra acestei chestiuni.

În fine, o altă proprietate a abaterii medii este faptul că valoarea ei este independentă faţă de schimbarea originii; valorile ce compun o serie pot fi mărite sau micşorate cu aceeaşi cantitate, valoarea abaterii tip va rămâne neschimbată.

Faptul poate să fie explicat şi fără o demonstraţie matematică. Într-adevăr, mărind toate valorile unei serii cu aceeaşi cantitate, media aritmetică a acesteia se va mări cu această câtime (§ 47); abaterile faţă de medie rămânând aşadar constante, atât fluctuaţia, cât şi abaterea-tip - care nu depind decât de aceste abateri - nu vor fi influenţate de această modificare. Fluctuaţia înălţimii unor arbori este aceeaşi, fie că înălţimea este măsurată de la nivelul terenului, fie de la nivelul mării.

De altfel, o altă probă a acestui lucru o constituie procedeul de calcul al abaterii tip expus în paragraful precedent. Originea *a* este arbitrară, însă oricum am fixa-o, calculul ne conduce la aceeaşi valoare a fluctuaţiei.

69. *Quartilele*. Abaterea medie şi abaterea-tip prezintă însă unele neajunsuri, pe care de altfel le-am întâlnit şi-n legătură cu media aritmetică. Explicaţia acestei coincidenţe este uşor de găsit: abaterile considerate au prin însăşi definiţia lor structura unei medii aritmetice, după cum ne putem uşor încredinţa comparând formulele (23) şi (24) cu (2).

Din această cauză, aceste două procedee de măsură a dispersiei nu pot să fie aplicate distribuţiilor deschise şi, în al doilea rând, valorile la care ne conduc sunt prea mult influenţate de elementele excepţionale pe care hazardul ni le poate introduce în distribuţii.

Pentru a se evita aceste obstacole, a fost nevoie de crearea unui al treilea mod de măsurare a dispersiei unei serii. Acest nou procedeu necesită calcularea în prealabil a două valori depinzând de seria respectivă şi cunoscute sub numele de *quartile*. Calculul quartilelor are la bază acelaşi principiu care ne-a condus la noţiunea de mediană. După cum ştim, mediana este valoarea care împarte în două părţi egal de numeroase ansamblul elementelor unei serii ordonat după mărime. Quartilele vor fi definite în acelaşi mod, cu simpla diferenţă că aceste valori împreună cu mediana vor împărţi acest ansamblu în patru părţi conţinând acelaşi număr de elemente.

Ca valoare, unul din quartile este mai mic decât mediana, iar altul mai mare decât ea; de unde şi numele de quartilul inferior şi quartilul superior.

În cazul când seria este formată din valori individuale, atunci quartilul inferior Q_1 şi cel superior Q_3 vor fi calculate după cum urmează:

1. *Seria conţine 4n elemente*. Însemnând prin l_k elementul de rangul k în ordinea de mărime crescătoare, vom avea:

$$Q_1 = \frac{l_n + l_{n+1}}{2} \qquad Q_3 = \frac{l_{3n} + l_{3n+1}}{2}$$

2. *Seria conține* $4n + 1$ *elemente*:

$$Q_1 = \frac{I_n + I_{n+1}}{2} \qquad Q_3 = \frac{I_{3n+1} + I_{3n+2}}{2}$$

3. *Seria conține* $4n + 2$ *elemente*:

$$Q_1 = I_{n+1} \qquad Q_3 = I_{3n+2}$$

4. *Seria conține* $4n + 3$ *elemente*:

$$Q_1 = I_{n+1} \qquad Q_3 = I_{3n+3}$$

Referindu-ne pentru o aplicație a acestor formule, la distribuția din tabelul 45 vom avea:

$$Q_1 = I_4 = 24{,}4 \qquad Q_3 = I_{12} = 26{,}5$$

care sunt quartilele seriei în timp corespunzătoare numărului de morți la mia de locuitori în România pentru anii 1901-1915.

Aceleași norme ne vor conduce pentru calcularea quartilelor unei distribuții de variabilă discontinuă. Astfel, să considerăm distribuția numărului de celule fermenți pe o suprafață divizată în milimetri pătrați (tabelul 60).

Cum 400 este egal cu 4×100, vom avea:

$$Q_1 = \frac{I_{100} + I_{101}}{2} = 3$$

$$Q_2 = \frac{I_{200} + I_{201}}{2} = 4 \quad (= Me)$$

$$Q_3 = \frac{I_{300} + I_{301}}{2} = 6.$$

În cazul distribuțiilor de variabilă continuă, considerații asemănătoare celor pentru cazul medianei (§ 51) - și pentru care motiv nu le mai repetăm - ne conduc să adoptăm pentru quartile valoarea elementelor de rangul $\frac{N}{4}$ și $\frac{3N}{4}$, indiferent de valoarea pe care o are N. Adică:

$$Q_1 = x_k + \frac{N - 4N_{k-1}}{4n_k} \cdot d$$

$$Q_3 = x_j + \frac{3N - 4N_{j-1}}{4n_j} \cdot d \qquad (28)$$

unde x și n sunt respectiv limitele inferioare și frecvențele intervalelor ce cuprind quartilele, iar N_{k-1}, N_{j-1}, frecvențele cumulate ale intervalelor ce preced imediat pe primele.

Aplicând aceste formule pentru calculul quartilelor distribuției în tabelul 48, obținem:

$$Q_1 = 117{,}5 + \frac{416 - 320}{224} \cdot 5 = 119{,}64,$$

$$Q_3 = 132{,}5 + \frac{1248 - 1200}{180} \cdot 5 = 133{,}83.$$

Ținând seamă și de valoarea mediană a aceleiași distribuții (§ 51), seria considerată se poate prezenta sub forma mai condensată din tabelul 61. Se vede din acest tabel care sunt intervalele care cuprind în ordinea lor de mărime câte 25% din cazurile examinate. Avem astfel o imagine mai comodă a distribuției respective.

TABELUL 60

Distribuția numărului de celule fermenți pe suprafețe egale cu 1 mm²

Numărul celulelor	Frecvențe	Frecvențe cumulate
0	0	0
1	20	20
2	43	63
3	**53**	**116**
4	**86**	**202**
5	70	272
6	**54**	**326**
7	37	363
8	18	381
9	10	391
10	5	396
11	2	398
12	2	400

Datele obținute de «Student»[1] sunt reproduse după R. A. Fisher, ***Statistical Methods for Research Workers****, London, Oliver & Boyd, 1928, p. 56.*

[1] *Pseudonimul unuia din cei mai însemnați cercetători în statistică matematică modernă, al cărui nume adevărat n-a fost încă divulgat. Lucrările sale se găsesc mai ales în* ***Biometrika****.*

Scanned with OKEN Scanner

TABELUL 61

Distribuția unui grup de adulți în raport cu greutatea

Greutatea în livre	Frecvențe absolute	Frecvențe relative
Sub 119,64	104	25
119,64 - 126,69	104	25
126,69 - 133,83	104	25
Peste 133,83	104	25

Tabelul 31

70. *Decilele*. Tot pentru o cunoaștere mai profundă a dispersiei s-au introdus niște valori analoage cu quartilele, numite *decile.* Decilele sunt valori care despart ansamblul ordonat al elementelor unei serii în zece grupuri egal de numeroase. Aceste valori sunt mai ales întrebuințate în cazul când distribuția este făcută în raport cu un număr prea mare de intervale, în scopul de a da acesteia din urmă o formă mai regulată.

Vom avea în total 9 decile, din care cel de-al cincilea se confundă cu mediana. Le vom însemna cu D_1, D_2,..., D_9. Formula care permite calculul lor este asemănătoare cu cele care ne dau valoarea medianei și a quartilelor:

$$D_r = x_i + \frac{rN - 10N_{i-1}}{10n_i} \tag{29}$$

unde D_r este cel de-al r-lea decil, egal deci cu elementul de rangul $\frac{rN}{10}$.

TABELUL 62

Distribuție de salarii în raport cu intervalele decile

Salariile în dolari	Frecvențe relative (%)
Sub 0,894	10
0,894-1,119	10
1,119-1,208	10
1,208-1,380	10
1,380-1,491	10
1,491-1,740	10
1,740-1,986	10
1,986-2,354	10
2,354-2,970	10
Peste 2,970	10

Tabelul 50

Scanned with OKEN Scanner

Valorile celor 9 decile ale distribuției de salarii din tabelul 50 sunt cuprinse în următoarele intervale: 0,85-0,94; 1,05-1,14; 1,15-1,24; 1,35-1,44; 1,45-1,54; 1,65-1,74; 1,95-2,04; 2,35-2,44; 2,95-3,04 și vor fi:

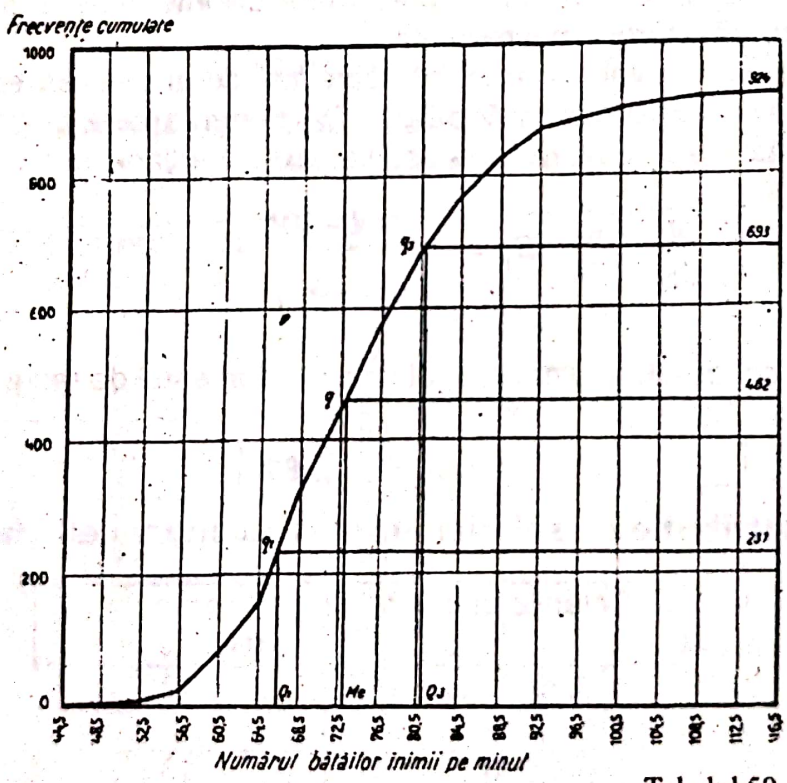
$$D_1 = 0{,}894 \quad D_2 = 1{,}119 \quad D_3 = 1{,}208 \quad D_4 = 1{,}380$$

$$D_5 = Me = 1{,}491$$

$$D_6 = 1{,}740 \quad D_7 = 1{,}986 \quad D_8 = 2{,}354 \quad D_9 = 2{,}970$$ [1].

Cu ajutorul acestor date, distribuția considerată se poate prezenta sub o formă mai strânsă, așa cum se vede în tabelul 62.

71. *Determinarea grafică a quartilelor și decilelor.* La fel ca și valoarea mediană, quartilele și decilele pot fi determinate cu o oarecare aproximație în mod grafic. În acest scop, se urmează procedeul cunoscut (§ 52), adică se recurge la imaginea grafică a distribuției sub forma cumulativă.



Tabelul 59

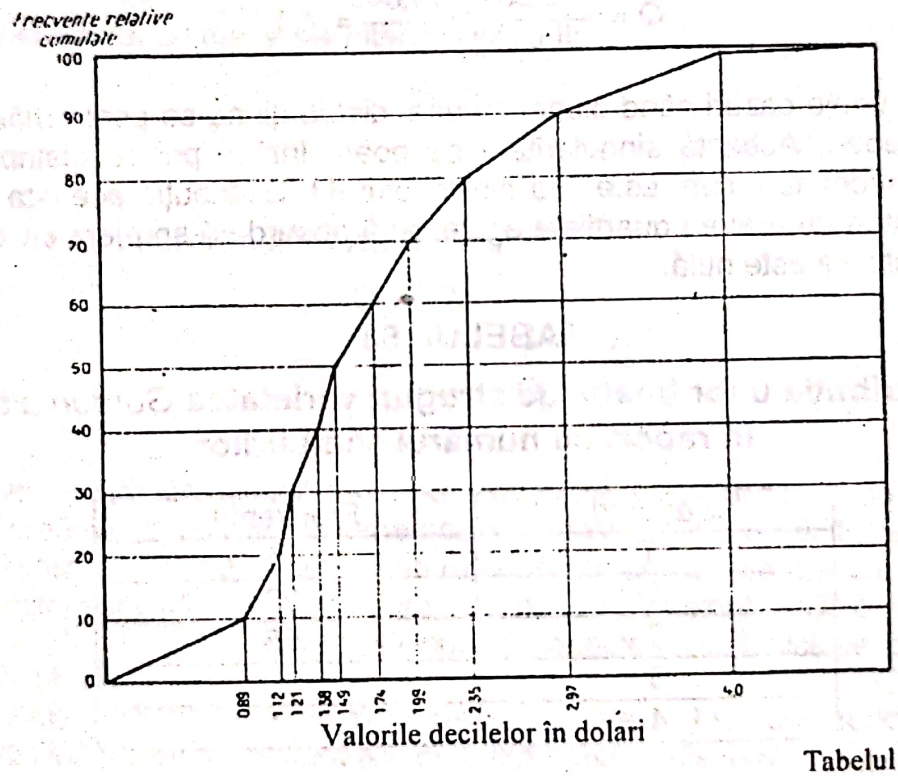
Graficul 31. Determinarea grafică prin intermediul ogivei a quartilelor și a medianei.

[1] *Din cauza formei incomode pentru tipar a distribuției corespunzătoare, nu redăm aceste calcule in extenso. Efectuarea lor nu prezintă nici o dificultate. Trebuie totuși să se aibă în vedere că, în conformitate cu considerațiile făcute mai sus (§18), limitele inferioare ale intervalelor trebuie deduse din cele din tabelul 50, prin diminuarea acestora cu 0,005. Astfel obținem limitele inferioare efective.*

Scanned with OKEN Scanner

Să ne fixăm ideile asupra distribuției din tabelul 59. Ogiva corespunzătoare acestei distribuții este cea din graficul 31. Împărțind ultima ordonată a ogivei în 4 părți egale și ducând prin punctele de diviziune paralele cu axa absciselor, obținem, prin intersecția lor cu ogiva, 3 puncte q_1, q_2, q_3, ale căror abscise ne dau valorile quartilelor și a medianei. Întrebuințând același procedeu, împărțind însă ultima ordonată a ogivei în 10 părți egale, obținem decilele.

În afară de faptul că această interpretare grafică a medianei, quartilelor și decilelor ne permite o bună evaluare rapidă a lor, ogiva mai poate servi pentru o reprezentare grafică mai schematică. Dispunem astfel de un procedeu foarte util pentru comparația grafică a distribuțiilor heterograde. Bineînțeles că în acest caz se calculează de obicei în prealabil valorile decilelor, iar ogiva se construiește pe baza acestora. Graficul 32 este obținut pe această cale; ogiva reprezintă grafic distribuția de salarii din tabelul 50, considerată însă sub forma din tabelul 62.



Tabelul 62

Graficul 32. Ogivă construită pe baza decilelor.

72. *Oscilația interquartilă.* Revenind acum la măsura dispersiei, observăm că 50% din elementele unei serii sunt cuprinse în intervalul mărginit de cele două quartile, și anume: 25% între Q_1 și *Me*, iar 25% între *Me* și Q_3. Cu cât valorile quartilelor vor fi mai apropiate de mediană, cu atât seria va prezenta o dispersie mai mică. Așadar, mărimea diferențelor între quartile și mediană constituie un element care variază cu dispersia seriei și care, prin urmare, poate fi considerat ca o măsură a acesteia.

Scanned with OKEN Scanner

În cazul unei distribuţii simetrice, quartilele se vor găsi în mod teoretic la aceeaşi distanţă de mediană. Intervalele Q_3 - Me, Me - Q_1 nu vor mai avea însă valori egale în cazul distribuţiilor asimetrice. Din această cauză, se adoptă ca măsură a dispersiei media aritmetică a acestor două intervale:

$$Q = \frac{(Q_3 - Me) + (Me - Q_1)}{2} = \frac{Q_3 - Q_1}{2}. \qquad (30)$$

Vom denumi această mărime cu termenul de *oscilaţie interquartilă*[1]. Ea se poate calcula chiar pentru distribuţiile deschise şi reprezintă de fapt *jumătatea oscilaţiei distribuţiei compusă numai din acele 50% din elemente care se grupează în jurul medianei respective.*

Vom avea pentru distribuţia din tabelul 31:

$$Q = \frac{133{,}83 - 119{,}64}{2} = 7{,}1.$$

Sunt unele cazuri când dispersia unei distribuţii nu se poate măsura prin acest procedeu. Această singularitate se poate întâlni printre distribuţiile de variabilă discontinuă, cum este cea din tabelul 63. Distribuţia aceasta prezintă particularitatea de a avea quartilele egale; ar fi absurd să spunem că dispersia corespunzătoare este nulă.

TABELUL 63

Distribuţia unor boabe de struguri varietatea *Gordon alb*, în raport cu numărul sâmburilor

Sâmburi	Frecvenţe
0	1
1	23
2	55
3	20
4	1

Distribuţie obţinută de dra M. Vestemeanu, absolventă a Şcolii de Statistică.

73. *Coeficienţi de dispersie.* Cele trei moduri de măsurare a dispersiei - aşa cum au fost examinate în paragrafele precedente - nu ne permit o evaluare a acesteia, independentă de ordinul de mărime al elementelor seriei. Or, ceea ce interesează mai mult este mai degrabă valoarea relativă a dispersiei, decât măsura ei absolută.

[1] *Această măsură corespunde valorii din calculul probabilităţilor numită* écart probable. *Termenul englezesc corespunzător este* interquartile range.

Scanned with OKEN Scanner

Astfel, grupurile de valori:

8, 10, 11,
998, 1000, 1001

au aceeaşi dispersie absolută, oricare ar fi procedeul de măsurare a acesteia. Este însă mai logic să asimilăm, în ceea ce priveşte dispersia, primul grup cu grupul: 800, 1000, 1100, decât cu cel menţionat.

Se obţine o măsură a dispersiei relative, dacă se raportează mărimile introduse în acest scop la valoarea ce reprezintă ordinul de mărime al elementelor unei distribuţii, adică la valoarea tipică a acesteia. Pentru această reducere, vom recurge la valoarea tipică corespunzătoare măsurii considerate.

Se obţine astfel *coeficientul mediu de dispersie,* dacă se calculează unul din raporturile $\frac{e_{Me}}{Me}$ sau $\frac{e_M}{M}$.

Coeficientul de dispersie interquartil va fi:

$$\frac{\frac{Q_3 - Q_1}{2}}{Me}$$

sau ceea ce este aproximativ acelaşi lucru:

$$q = \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}. \qquad (31)$$

Acest coeficient prezintă unul din avantajele pe care ar fi de dorit să le aibă toţi coeficienţii statistici, anume acela de a varia între două limite fixe, - 1 şi + 1, de preferinţă. Păcat însă că, pentru coeficientul considerat, cazurile în care aceste limite sunt atinse ($Q_1 = 0$ sau $Q_3 = 0$) nu prezintă nici o semnificaţie particulară, după cum se întâmplă cu alţi coeficienţi statistici pe care îi vom întâlni mai târziu[1].

Cel mai întrebuinţat coeficient de dispersie este însă cel introdus de prof. K. PEARSON[2]. Acest coeficient are ca bază măsura dispersiei prin abaterea tip şi poartă numele de *coeficient de variabilitate*:

$$C.V. = \frac{\sigma}{M} \cdot 100. \qquad (32)$$

Importanţa acestui coeficient va fi mai bine evidenţiată cu prilejul discuţiei asupra comparaţiei distribuţiilor (Cap.IX).

---

[1] *Vezi Cap.X, XI, XIV.*

[2] *Karl Pearson,* ***Regression, Heredity and Panmixia,*** *Philosophical Transactions of Royal Society, A, vol. CLXXXVII, 1896, p. 253.*

Scanned with OKEN Scanner

Deocamdată vom insista asupra necesității ca pentru o comparație de dispersii să întrebuințăm același coeficient și pentru o serie și pentru cealaltă. Aceasta, din cauză că, pentru aceeași serie chiar, cele trei moduri de măsurare a dispersiei ne pot prea bine conduce la rezultate diferite (tabelul 64).

Nu trebuie văzut în acest fapt un punct slab al metodei statisticii; dispersia este o noțiune elastică și este natural ca, adoptând puncte de vedere diferite pentru măsura ei, să obținem și rezultate diferite.

TABELUL 64

Comparația între diferitele măsuri ale dispersiei

Modul de măsură	Dispersia	
	absolută	relativă
e_M	8,82	0,07
σ	11,20	0,09
Q	7,10	0,06

Tabelul 31

74. *Asimetrie. Coeficienți pearsonieni.* În afară de dispersie, forma unei distribuții poate să fie privită și din alte puncte de vedere. Mai sunt, în afară de dispersie, încă două calități ale distribuțiilor. Acestea se pun mai ușor în evidență prin reprezentarea grafică și pot constitui elemente de diferențiere a distribuțiilor.

După cum am văzut cu prilejul formei distribuțiilor (§ 34), o astfel de serie statistică poate fi simetrică sau asimetrică. Chiar două distribuții asimetrice se pot deosebi prin *gradul* de asimetrie ce-l prezintă fiecare, cât și prin felul acestei asimetrii.

Astfel, curba A din graficul 33 este mai asimetrică decât B. În afară de aceasta, prima curbă este *pozitiv asimetrică*, iar B, *negativ asimetrică*, dacă se convine să se considere asimetria pozitivă sau negativă după cum coada lungă a distribuției este de partea valorilor mari ale seriei sau de partea celor mici.

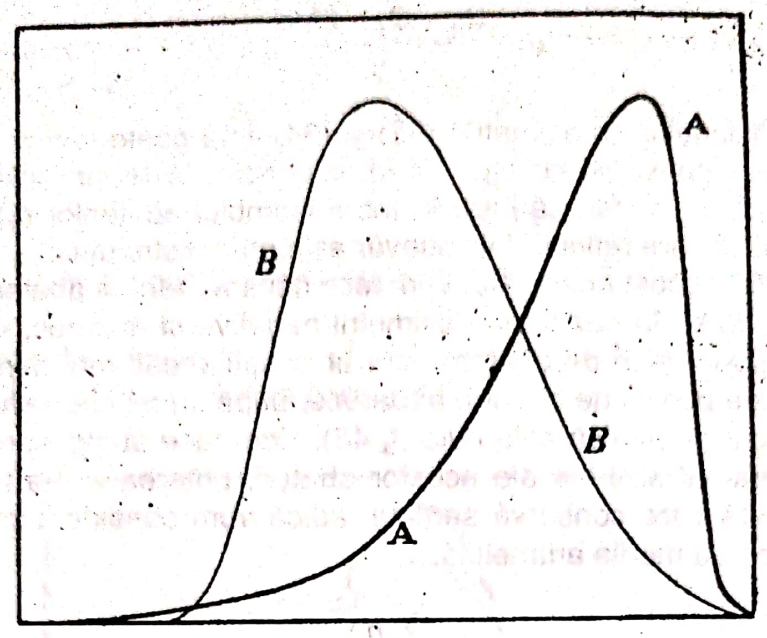
Evident că asimetria va fi cu atât mai pronunțată, cu cât diferența între media aritmetică și modul seriei va fi mai mare, distribuțiile simetrice caracterizându-se prin egalitatea acestor două valori tipice.

Este iarăși evident că distribuțiile reprezentate grafic prin curbele de frecvență din graficul 34 trebuie considerate la fel de asimetrice. Cu toate acestea, diferența între media aritmetică și mod nu este aceeași pentru ambele cazuri, ceea ce se explică prin faptul că cele două distribuții considerate n-au aceeași dispersie.

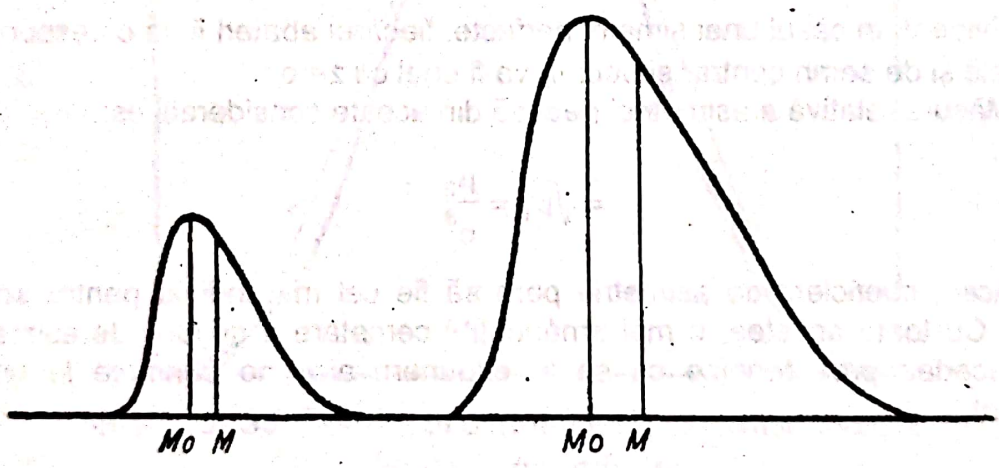
Așadar, pentru a avea o justă măsură a asimetriei prin diferența între media aritmetică și mod, trebuie să raportăm această diferență la mărimea absolută a dispersiei; obținem astfel coeficientul de asimetrie pearsonian:

$$sk = \frac{M - Mo}{\sigma} \quad [1] \qquad (33)$$

[1] *Notația acestui coeficient introdus de prof. Pearson este legată de termenul* skewness, *întrebuințat de d-sa pentru denumirea asimetriei. K. Pearson,* ***Skew Variation in Homogeneous Material****, Phil.Trans. Roy.Soc., A. vol. CLXXXVI, 1895, p.343.*



Graficul 33. Distribuţii prezentând grade de asimetrie diferite.



Graficul 34. Distribuţii prezentând acelaşi grad de asimetrie şi dispersii diferite.

Un alt procedeu simplu de a măsura asimetria unei distribuţii este următorul.

Am văzut, mai sus, că în intervalele *Me* - Q_1, Q_3 - *Me* se găsesc acelaşi număr de elemente ale seriei respective, anume câte 25%. Egalitatea acestor două intervale ar putea fi luată ca un indiciu al simetriei distribuţiei, diferenţa lor deci, ca o măsură a asimetriei absolute. Asimetria relativă va fi măsurată prin raportarea acestei diferenţe la dispersia măsurată din acelaşi punct de vedere, adică la oscilaţia interquartilă. Obţinem astfel *coeficientul de asimetrie interquartil:*

Scanned with OKEN Scanner

$$\alpha = \frac{Q_3 + Q_1 - 2Me}{Q}. \qquad (34)$$

Însă coeficienţilor de asimetrie (33) şi (34) li se poate imputa faptul că şi unul, şi altul sunt construiţi cu ajutorul unor puncte de reper izolate - modul, media, quartilele - şi nu fac să intre în joc ansamblul abaterilor faţă de media aritmetică, singurul care reflectă într-adevăr asimetria distribuţiei.

Considerând acest ansamblu, vom face constatarea că abaterile negative vor fi mai numeroase în cazul unei asimetrii negative şi reciproc. O rezultantă directă a celor două feluri de abateri, care ar constitui cea mai bună măsură a asimetriei, nu este posibil de obţinut, deoarece, după cum ştim, suma abaterilor în raport cu media aritmetică este nulă (§ 48). Vom face atunci suma algebrică a puterilor de ordinul al 3-lea ale acestor abateri, puterea a 3-a fiind puterea imediat superioară care conservă semnul, adică vom considera momentul de ordinul 3 în raport cu media aritmetică:

$$\mu_3 = \frac{\sum n \xi^3}{N}.$$

Evident, în cazul unei simetrii perfecte, fiecărei abateri îi va corespunde o alta egală şi de semn contrar şi deci μ_3 va fi egal cu zero.

Măsura relativă a asimetriei dedusă din aceste consideraţii este

$$\gamma = \sqrt{\beta_1} = \frac{\mu_3}{\sigma^3}\ ^{1}. \qquad (35)$$

Acest coeficient de asimetrie pare să fie cel mai indicat pentru scopul propus. Cu toate acestea, o mai amănunţită cercetare a noţiunii de asimetrie, prin procedee prea tehnice ca să le expunem aici, ne conduce la un alt coeficient:

$$sk = \frac{\sqrt{\beta_1}(\beta_1 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} \qquad (36)$$

care nu este decât o evaluare mai exactă a coeficientului (33). Ştim, într-adevăr, că acesta din urmă necesită calculul prealabil al modului care nu este o valoare prea stabilă[2].

Formula (36) are însă alt neajuns: este complicată şi lipsită de caracterul intuitiv al celorlalte.

---

[1] *A se vedea mai departe formulele (37).*

[2] *K. Pearson,* ***On the General Theory of Skew Correlation and Non-linear Regression****, Drapers' Company Research Memoirs, Biometric Series II.*

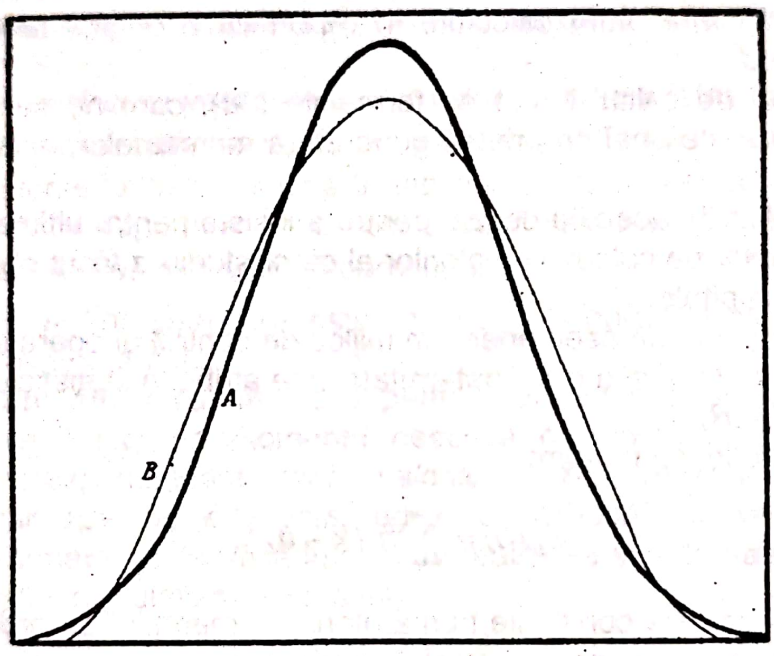
Scanned with OKEN Scanner

De altfel, ea depinde de cantităţile β_1 şi β_2, numite *coeficienţi pearsonieni* şi definiţi de relaţiile:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \qquad \beta_2 = \frac{\mu_4}{\mu_2^2} \tag{37}$$

unde μ_3, μ_4 sunt respectiv momentele de ordinul 3 şi 4 în raport cu media aritmetică.

Procedeele rapide de calcul ale acestor momente, precum şi câteva aplicaţii ale lor vor face obiectul paragrafului 76.



Graficul 35. Distribuţii simetrice de aceeaşi dispersie, dar de boltiri diferite.

75. *Boltire.* Ne vom ocupa acum de ultimul aspect intuitiv al formei unei distribuţii: de *boltire*[1]. Graficul 35 ilustrează diferenţa care poate exista din punct de vedere al boltirii între două curbe de frecvenţe simetrice şi de aceeaşi dispersie.

Măsura relativă a boltirii este dată de coeficientul pearsonian de ordinul 2, β_2. Comparaţia se face în raport cu boltirea curbei normale, când $\beta_2 = 3$. Dacă $\beta_2 > 3$, curba este mai ascuţită spre vârf decât curba normală, dacă $\beta_2 < 3$, mai turtită. Termenul tehnic este acela de distribuţie *leptokurtică*, în cazul când $\beta_2 > 3$ şi *platykurtică*, când $\beta_2 < 3$ [2].

---

[1] Voussure *în franceză;* kurtosis *în engleză.*

[2] *Din greacă:* leptos = *subţire,* platys = *lat.*

76. *Calculul momentelor în raport cu originea. Proba lui Charlier.* Coeficienţii statistici introduşi până în prezent necesită, după cum am văzut, calculul momentelor de ordinul 3 şi 4 în raport cu media aritmetică. Aceasta se poate face prin procedee simplificate asemănătoare celor întrebuinţate pentru calcularea mediei şi fluctuaţiei. Formulele care stau la baza acestor procedee sunt următoarele:

$$\mu_3 = \left[m_3(a) - 3m_2(a)\, m_1(a) + 2\overline{m_1(a)}^3\right] \times d^3$$

$$\mu_4 = \left[m_4(a) - 4m_3(a)\, m_1(a) + 6m_2(a)\, \overline{m_1(a)}^2 - 3m_1(a)^4\right] \times d^4 \qquad (38)$$

unde momentele $m(a)$ sunt calculate în raport cu o origine arbitrară a şi în unităţi intervale d.

Procedeul de calcul bazat pe formulele (38), care nu sunt decât nişte cazuri particulare ale unei proprietăţi generale a momentelor, este ilustrat prin tabelul 65.

Vom profita de această ocazie pentru a insista pentru ultima oară asupra tuturor procedeelor de calcul ale valorilor al căror studiu a făcut obiectul acestor două din urmă capitole.

Vom menţiona, de asemenea, un mijloc de control al operaţiilor relative la calcularea momentelor şi a cărui paternitate este atribuită statisticianului suedez C. V. L. CHARLIER.

În acest scop, să punem:

$$s_p(a) = \sum n\,(x-a)^p . \qquad (39)$$

Această expresie constituie numărătorul momentului de ordinul p şi este prima valoare care se obţine prin totalul unei coloane, aşa după cum am văzut în exemplele precedente, atunci când se calculează acest moment. Cu alte cuvinte:

$$m_p(a) = \frac{s_p(a)}{N} .$$

Să presupunem acum că vrem să calculăm momentul de ordinul 4. Relaţiile (38) ne arată că în acest caz suntem obligaţi, dacă întrebuinţăm procedeul simplificat, să calculăm toate momentele până la ordinul considerat. Pentru aplicarea probei lui Charlier, avem nevoie de un calcul auxiliar, şi anume: vom calcula momentul de ordinul 4, în raport cu o valoare care să difere de originea aleasă cu mărimea unui interval, adică se va obţine $s_4(a\text{-}d)$ (coloana 10, tabelul 65). Apoi, se va căuta să se vadă dacă valorile $s_0(a)$, $s_1(a)$, $s_2(a)$, $s_3(a)$, $s_4(a)$ verifică relaţia:

Scanned with OKEN Scanner

$$s_4(a-d) = s_4(a) - 4ds_3(a) + 6d^2s_2(a) - 4d^3s_1(a) + d^4s_0(a) \qquad (40)$$

care devine, în cazul când se adoptă lungimea intervalelor ca unitate,

$$s_4(a - 1) = s_4(a) - 4s_3(a) + 6s_2(a) - 4s_1(a) + s_0(a) .$$

Bineînțeles că pentru cazul momentelor de ordin inferior vom întrebuința după împrejurări:

$$s_3(a-d) = s_3(a) - 3ds_2(a) + 3d^2s_1(a) - d^3s_0(a)$$

$$s_2(a-d) = s_2(a) - 2ds_1(a) + d^2s_0(a) . \qquad (40 \text{ bis})$$

Proba lui Charlier constă din aplicarea acestor relații; ea nu ne dă decât indicații de ordin negativ, însă aceasta este tocmai partea cea mai importantă a controlului calculelor.

Tot ca un mijloc de verificare a calculelor ne putem folosi de tablele care ne dau pe $n\alpha^4$, pentru diferite valori ale lui n și α [1]. După ce am obținut prin calcul direct valorile lui $n\alpha^4$ (coloana 9, tabelul 65), le putem confrunta cu cele date de table; orice nepotrivire între aceste două valori ne obligă să controlăm calculul pentru valoarea corespunzătoare lui α.

77. *Corecția Sheppard.* Mai este de făcut o singură observație în ceea ce privește calculul momentelor.

După cum am văzut mai sus pentru calculul mediei aritmetice și al fluctuației, orice calcul de momente necesită o transformare a distribuției heterograde corespunzătoare: *toate valorile dintr-un interval sunt înlocuite printr-un număr egal de elemente, egale cu mijlocul intervalului.* Or, prin această transformare, se falsifică întrucâtva realitatea și este natural ca valorile momentelor astfel calculate să se resimtă.

Corecțiile care trebuie aduse momentelor brute ν pentru a obține momentele ajustate μ sunt cunoscute sub numele de *corecțiile lui SHEPPARD*[2]. Acestea sunt, în ceea ce ne privește, exprimate prin relațiile:

$$\mu_2 = \nu_2 - \frac{d^2}{12}$$

$$\mu_4 = \nu_4 - \frac{d^2}{2}\nu_2 + \frac{7d^4}{240}. \qquad (41)$$

---

[1] K. Pearson, ***Tables for Statisticians and Biometricians***, *Cambridge University Press, 1930, Part I, 3rd edition, p. 102-112.*

[2] *W. F. Sheppard,* ***On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equidistant Divisions of a Scale****, Proc. Lon. Math. Soc., vol. XXIX, p. 353.*

Scanned with OKEN Scanner

Media aritmetică, după cum știm, nu este afectată de transformarea menționată (§ 43). Un calcul mai amănunțit ne arată că nici momentul de ordinul 3, în cazul când este calculat în raport cu media aritmetică, nu prezintă corecție din acest punct de vedere.

Formulele (41) nu se pot aplica în următoarele cazuri:

1. Când avem de-a face cu o distribuție de valori individuale, căci atunci dispare motivul care ne-a condus la corecție.
2. Când intervalele marginale ale distribuției nu prezintă frecvențe relativ mici, cum se întâmplă, de pildă, în cazul distribuțiilor în U și J. Corecțiile de aplicat momentelor acestor distribuții sunt prea complicate ca să le mai menționăm aici; din fericire, astfel de distribuții se întâlnesc destul de rar[1].

[1] *K. Pearson, **Tables for Statisticians and Biometricians**, Cambridge University Press, 1931, Part II, p. CLXXXVII.*

Scanned with OKEN Scanner

TABELUL 65

Calculul valorilor și coeficienților statistici pentru distribuția bărbaților căsătoriți, în raport cu vârsta

Vârsta	Mijlocul intervalelor x_i	Frecvențe n_i	Frecvențe cumulate N_i	Abateri în intervale α_i	Produsele[1] $n_i\alpha_i^1$	$n_i\alpha_i^2$	$n_i\alpha_i^3$	$n_i\alpha_i^4$	$n_i(\alpha_i-1)^4$	Abateri față de medie δ_i	Produse $n_i\delta_i$	Abateri față de mediană Δ_i	Produse $n_i\Delta_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
15 - 20	17,5	2	2	- 5	- 10	50	- 250	1250	2592	-27,85	- 55,70	-26,72	- 53,44
20 - 25	22,5	38	40	- 4	- 152	608	- 2432	9728	23750	-22,85	- 868,30	-21,72	- 825,36
25 - 30	27,5	96	136	- 3	- 288	864	- 2592	7776	24576	-17,85	-1713,60	-16,72	-1605,12
30 - 35	32,5	123	259	- 2	- 246	492	- 984	1968	9963	-12,85	-1580,55	-11,72	-1441,56
35 - 40	37,5	132	391	- 1	- 132	132	- 132	132	2112	- 7,85	-1036,20	- 6,72	- 887,04
40 - 45	42,5	129	520	0	**- 828**	-	**- 6390**	-	129	- 2,85	- 367,65	- 1,72	- 221,88
45 - 50	47,5	123	643	+ 1	+ 123	123	+ 123	123	-	+ 2,15	+ 264,45	+ 3,28	+ 403,44
50 - 55	52,5	110	753	+ 2	+ 220	440	+ 880	1760	110	+ 7,15	+ 786,50	+ 8,28	+ 910,80
55 - 60	57,5	88	841	+ 3	+ 264	792	+ 2376	7128	1408	+12,15	+1069,20	+13,28	+1168,64
60 - 65	62,5	70	911	+ 4	+ 280	1120	+ 4480	17920	5670	+17,15	+1200,50	+18,28	+1279,60
65 - 70	67,5	46	957	+ 5	+ 230	1150	+ 5750	28750	11776	+22,15	+1018,90	+23,28	+1070,88
70 - 75	72,5	26	983	+ 6	+ 156	936	+ 5616	33696	16250	+27,15	+ 705,90	+28,28	+ 735,28
75 - 80	77,5	12	995	+ 7	+ 84	588	+ 4116	28812	15552	+32,15	+ 385,80	+33,28	+ 399,36
80 - 85	82,5	4	999	+ 8	+ 32	256	+ 2048	16384	9604	+37,15	+ 148,60	+38,28	+ 153,12
85 - 90	87,5	1	1000	+ 9	+ 9	81	+ 729	6561	4096	+42,15	+ 42,15	+43,28	+ 43,28
		1000			**+1398**	**7632**	**+26118**	**161988**	**127588**		**-5622,00**		**-5034,40**
											+5622,00		**+6164,40**

Tabelul 12

[1] *Fiecare din coloanele 7, 8, 9 sunt obținute prin înmulțirea cu α_i a valorilor coloanei ce o precede. Prin acest procedeu evităm calcularea în prealabil a puterilor α^2, α^3, α^4. Valorile coloanei (10) se pot obține fie direct, fie prin intermediul tablelor menționate (§ 76).*

Scanned with OKEN Scanner

Calcule referitoare la tabelul 65

Proba lui Charlier

$$s_0(a) = 1000 \quad s_1(a) = +570 \quad s_2(a) = 7632 \quad s_3(a) = +19728$$

$$s_4(a) = 161988 \quad s_4(a+1) = 127588$$

$$161988 - 4 \times 19728 + 6 \times 7632 - 4 \times 570 + 1000 = 127588$$

Calculul momentelor

$m_1(a) = 0{,}57$ $\quad M = 42{,}5 + 0{,}57 \times 5 = 45{,}35$

$m_2(a) = 7{,}632$ $\quad \nu_2 = \left(7{,}632 - 0{,}57^2\right) \times 5^2 = 7{,}3071 \times 25 = 182{,}6775$

$$\mu_2 = \left(7{,}632 - 0{,}57^2 - \frac{1}{12}\right) \times 5^2 = 7{,}22377 \times 25 = 180{,}5943$$

$m_3(a) = 19{,}728$ $\quad \nu_3 = \mu_3 = \left(19{,}728 - 3 \times 7{,}632 \times 0{,}57 + 2 \times 0{,}57^3\right) \times 125 =$

$$= 7{,}047666 \times 125 = 880{,}958250$$

$m_4(a) = 161{,}988$ $\quad \nu_4 = (161{,}988 - 4 \times 19{,}728 \times 0{,}57 + 6 \times 7{,}632 \times 0{,}57^2 - 3 \times$

$\times 0{,}57^4) \times 625 = 131{,}56930077 \times 625 = 82230{,}81298125$

$$\mu_4 = \left(131{,}56930077 - \frac{7{,}3071}{2} + \frac{7}{240}\right) \times 625 =$$

$$= 127.94491744 \times 625 = 79965{,}5734$$

Calculul altor valori tipice

$$Me = 40 + \frac{1000 - 782}{258} \times 5 = 44{,}22$$

$$Q_1 = 30 + \frac{1000 - 544}{492} \times 5 = 34{,}63$$

$$Q_3 = 50 + \frac{3000 - 2572}{440} \times 5 = 54{,}86$$

$$Mo = 37{,}5 + \frac{129 - 123}{264 - 129 - 123} \times \frac{5}{2} = 38{,}75 \qquad \text{Formula (11)}$$

$$Mo = 45{,}35 - 3(45{,}35 - 44{,}22) = 41{,}96 \qquad \text{Formula (12)}$$

Scanned with OKEN Scanner

Coeficienți de dispersie

$$e_M = \frac{11244}{1000} = 11{,}244 \qquad \frac{e_M}{M} = \frac{11{,}244}{45{,}35} = 0{,}248$$

$$e_{Me} = \frac{11198}{1000} = 11{,}198 \qquad \frac{e_{Me}}{Me} = \frac{11{,}198}{44{,}22} = 0{,}253$$

$$\sigma = \sqrt{7{,}2238} \times 5 = 13{,}438 \qquad C.V. = \frac{13{,}438}{45{,}35} \times 100 = 29{,}63$$

$$Q = \frac{54{,}86 - 34{,}63}{2} = 10{,}115 \qquad q = \frac{54{,}86 - 34{,}63}{54{,}86 + 34{,}63} = 0{,}226$$

Coeficienți pearsonieni

$$\beta_1 = \frac{49{,}669596}{376{,}956928} = 0{,}131864 \qquad \beta_2 = \frac{127{,}944917}{52{,}182853} = 2{,}452$$

Coeficienți de asimetrie

$$sk = \frac{M - Mo}{\sigma} = \frac{45{,}35 - 38{,}75}{13{,}438} = \frac{6{,}60}{13{,}438} = 0{,}4911$$

$$sk = \frac{45{,}35 - 41{,}96}{13{,}438} = \frac{3{,}39}{13{,}438} = 0{,}2523$$

$$\alpha = \frac{Q_1 + Q_3 - 2Me}{Q} = \frac{1.05}{10{,}115} = 0{,}1038$$

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \sqrt{0{,}131864} = 0{,}363$$

$$sk = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)} = \frac{1{,}979076}{4{,}937632} = 0{,}4018$$

Scanned with OKEN Scanner

Capitolul VI

Distribuția normală, ajustare

78. *Distribuția normală*. Dintre toate tipurile de distribuții ce se pot întâlni în practică, unul din cazurile cele mai importante îl constituie, fără îndoială, *distribuția normală*.

Această distribuție face parte din grupul de distribuții pentru care curba frecvențelor corespunzătoare se prezintă sub forma profilului unui clopot. Ea este, așadar, caracterizată în primul rând prin faptul că este simetrică. Toți coeficienții de asimetrie introduși în paragrafele precedente vor avea în acest caz valoarea zero. O a doua caracteristică a distribuției normale este faptul că valoarea coeficientului de boltire β_2 este egală cu 3.

În practică însă, nici coeficienții de asimetrie, nici β_2 nu vor avea decât în mod accidental valorile indicate de teorie. Cu toate acestea, abaterile lor față de aceste valori nu vor putea depăși, în majoritatea cazurilor, niște limite bine definite fără ca ipoteza normalității distribuției să fie pusă în discuție (despre aceste limite vom vorbi în capitolul următor).

Coeficienții de asimetrie și de boltire nu fac totuși altceva decât să ne dea o indicație de ordin negativ în ceea ce privește chestiunea dacă o distribuție este normală sau nu. Caracterizarea completă a ei nu se poate face decât prin ecuația curbei de frecvențe corespunzătoare:

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(x-M)^2}{2\sigma^2}} \qquad (42)$$

unde y este ordonata aferentă valorii variabilei independente x.

Celelalte notații ne sunt deja cunoscute, cu excepția lui e care este o constantă matematică, de aceeași structură ca și numărul π și egală cu 2,7182818....

Curba este cea reprezentată de graficul 36 și este cunoscută sub numele de curba (legea) lui LAPLACE-GAUSS, după numele celor doi matematicieni care s-au ocupat mai de aproape cu studiul ei.

Fenomene naturale care să se distribuie în mod neîndoios după legea normală sunt foarte rare. Totuși, dacă ne mulțumim cu o primă aproximare a lucrurilor, legea normală constituie o reprezentare minunată a unui foarte variat

Scanned with OKEN Scanner

ansamblu de fenomene. Distribuţii tipice normale nu se întâlnesc decât printre acelea de măsuri biometrice, deşi şi aici se pot găsi multe excepţii, cum este cazul clasic al greutăţii corporale.

În această direcţie, statisticianul QUÉTELET i-a supraevaluat aplicaţiile, dar faptul nu poate fi considerat ca o eroare din parte-i, deoarece el a rămas la studiul aproximat al distribuţiilor.

În afară de biometrie, legea normală îşi găseşte aplicaţii la studiul *erorilor întâmplătoare*, cum sunt erorile de observaţie, erorile statistice provenite din jocul hazardului, erorile de tragere la ţintă etc.

De aceea, această lege este cunoscută şi sub numele de legea erorilor sub care faţă a constituit preocupările de cercetare ale lui LAPLACE şi GAUSS.

QUÉTELET, făcând o legătură între aceste două domenii de aplicabilitate ale legii normale, ajunge la crearea noţiunii de **individ biometric mediu.** După interpretarea dată de QUÉTELET, natura trage la ţintă, ochind acest tip mediu; indivizii care se produc se vor distribui după legea erorilor, erori pe care natura le comite în această operaţie de ţintire.

Această teorie a lui QUÉTELET, deşi nu mai are azi decât prea puţine puncte de sprijin, a avut totuşi marele merit de a fi atras atenţia asupra unei regularităţi reale, provocând astfel o întreagă serie de cercetări asupra variabilităţii biometrice a speciilor.

Azi, după studii mai aprofundate, legea normală nu mai este considerată decât ca o interesantă aproximaţie a acestei variabilităţi, însă importanţa ei pentru metoda statistică, din cauza interpretării ce o dă legilor hazardului, creşte pe fiecare zi.

79. *Ajustarea negraduată a distribuţiei normale.* Operaţia prin care se caută ca pe o distribuţie heterogradă obţinută printr-o înregistrare statistică să se calcheze o curbă de frecvenţă, care, după cum ştim, reprezintă întrucâtva distribuţia ideală care s-ar obţine în cazul când nu ar interveni hazardul, poartă numele de *ajustare.*

Ajustarea unei curbe normale se face în două feluri, după cum distribuţia este continuă sau nu. Calculul este din cele mai simple şi se sprijină pe un şir de valori date de table speciale.

Să considerăm o distribuţie discontinuă (tabelul 66).

Vom calcula în primul rând valorile tipice şi coeficienţii statistici introduşi în capitolele precedente.

$$m_1 = -0{,}1556 \qquad m_2 = 2{,}24118166$$
$$m_3 = 0{,}02160494 \qquad m_4 = 16{,}606264$$

$$M = 9{,}8444 \qquad \mu_2 = 2{,}21697030 \qquad \sigma_1 = 1{,}4889$$
$$\gamma_1 = 0{,}308 \qquad \beta_2 = 3{,}442$$

[1] *Fără* corecţia Sheppard, *deoarece ne găsim în cazul unei distribuţii în raport cu valori individuale ale variabilei independente.*

Ultimii doi coeficienți au fost calculați în scopul de a ne face o idee de cât de mult se depărtează distribuția considerată de distribuția ideal normală; ceilalți, pentru că, după cum se vede din (42), intră în compoziția acestei formule. Pentru o ajustare de curbă de tip normal vom avea, așadar, nevoie *numai de calculul prealabil al mediei aritmetice și al abaterii-tip.*

Calculul propriu-zis al ajustării se face după cum urmează:

1. Se calculează în primul rând abaterile reduse (coloanele 2, 3, tabelul 66).

$$X = \frac{x - M}{\sigma}.$$

2. Utilizând Tablele I[1] de la sfârșitul cărții, obținem valorile:

$$z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

corespunzătoare acestor abateri (coloana 4).

TABELUL 66

Ajustarea curbei de tip normal pe o distribuție de flori de mac în raport cu numărul stigmatelor

Numărul stigmate-lor	Abateri $x - M$	Abateri relative $(x - M)/\sigma$	z	$\frac{N}{\sigma} \times z$	Frecvențe teoretice	Frecvențe statistice
(1)	(2)	(3)	(4)	(5)	(6)	(7)
4	- 5,8444	- 3,9253	0,0001799	0,27	0,3	0
5	- 4,8444	- 3,2536	0,0020056	3,06	3,0	1
6	- 3,8444	- 2,5820	0,0142321	21,68	21,7	12
7	- 2,8444	- 1,9104	0,0643289	97,99	98,0	91
8	- 1,8444	- 1,2388	0,1852131	282,13	282,1	295
9	- 0,8444	- 0,5671	0,3396815	517,43	517,4	550
10	+ 0,1556	+ 0,1045	0,3967650	604,38	604,4	619
11	+ 1,1556	+ 0,7761	0,2952021	449,67	449,7	418
12	+ 2,1556	+ 1,4478	0,1398771	213,07	213,1	195
13	+ 3,1556	+ 2,1194	0,0422202	64,31	64,3	54
14	+ 4,1556	+ 2,7911	0,0081151	12,36	12,4	25

[1] Sau ***Tables for Statisticians***, *Part I, p. 2.*

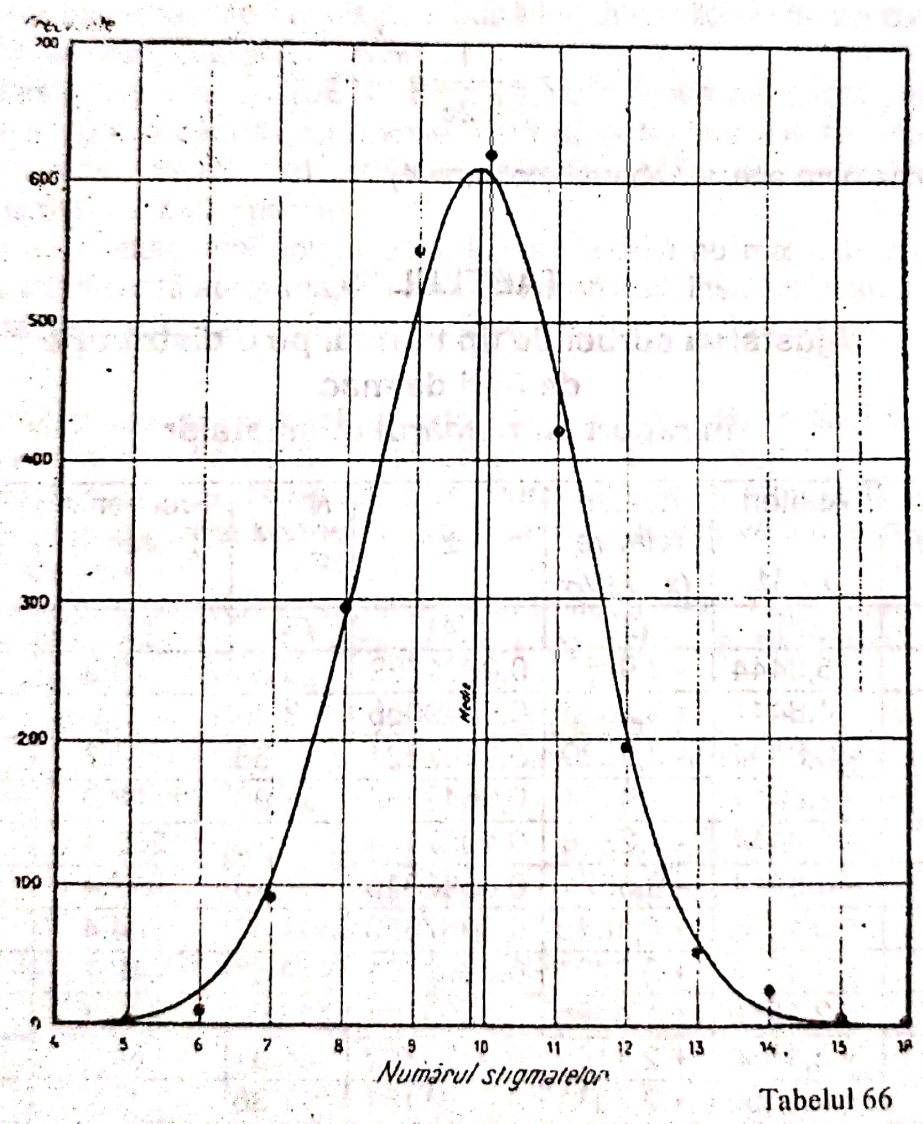
Scanned with OKEN Scanner

(continuare tab. 66)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
15	+ 5,1556	+ 3,4627	0,0009938	1,51	1,5	5
16	+ 6,1556	+ 4,1343	0,0000775	0,12	0,1	3
					2268,0	**2268**

*Distribuție reprodusă după E. Coquidé, **op. cit.**, p. 101.*

$$\frac{N}{\sigma} = 1523{,}2722 \qquad d = 1.$$



Tabelul 66

Graficul 36. Curba normală ajustată pe distribuția unor flori de mac în raport cu numărul stigmatelor.

Este de observat aici faptul că, pentru a găsi valorile lui z din table, vom utiliza valorile absolute ale abaterilor relative $\frac{x-M}{\sigma}$. De asemenea, trebuie observat că valorile lui *z* sunt obţinute prin *interpolare liniară*. Astfel, găsim în table:

$\frac{x-M}{\sigma}$	*z*
3,92	0,0001837
3,93	0,0001766

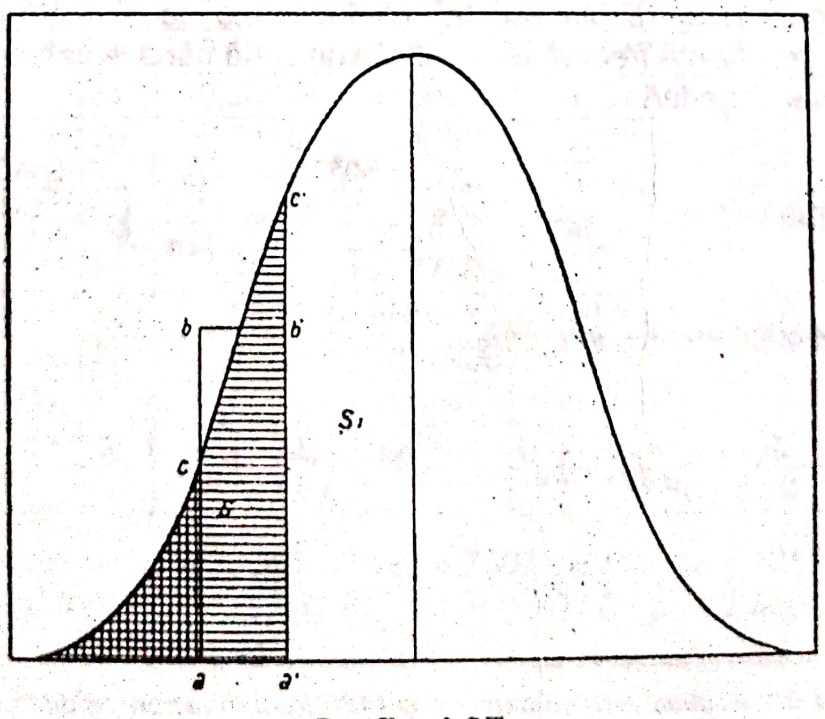
Diferenţa dintre valorile corespunzătoare lui *z* fiind 71, vom avea pentru abaterea relativă 3,9253, valoarea:

$$0{,}0001837 - \frac{71 \times 53}{100} = 0{,}0001799.$$

3. Înmulţim aceste valori cu $\frac{N}{\sigma} \times d$ pentru a obţine frecvenţele absolute ale curbei de frecvenţă ideală (coloana 5 şi 6).

Această operaţie se completează prin reprezentarea grafică a curbei de frecvenţe astfel obţinută (graficul 36).

80. *Ajustare graduată.* Ajustarea curbei de tip normal pentru o distribuţie de variabilă continuă diferă în parte de modul cum se face această operaţie în cazul unei distribuţii discontinue.



Graficul 37.

Să considerăm distribuţia lungimii unor tije de orz obţinută de BLARINGHEM şi reprodusă după *G. K. Constantinescu*, ***Tratat de zootehnie***, Bucureşti, 1930, p. 176. Vom avea prilejul să constatăm cu această ocazie cât de reprezentativă este curba normală pentru o distribuţie de talii.

Amănuntele calculului se văd în tabelul 67. Evident, şi în acest caz va trebui să calculăm în prealabil momentele distribuţiei considerate, cu singura deosebire că vom obţine valorile acestora după aplicarea corecţiei SHEPPARD.

În continuare, pentru a obţine frecvenţele teoretice aferente claselor distribuţiei noastre, este necesar să remarcăm că variabila independentă fiind continuă, frecvenţele vor fi reprezentate prin suprafeţe: cele statistice, prin dreptunghiuri *aa'bb'*, cele teoretice, prin suprafeţe mărginite de curba de frecvenţe, *aa'cc'* (graficul 37).

Or, suprafaţa *aca'c'* nu poate fi calculată prin mijloace elementare. Ea poate fi considerată ca diferenţa a două suprafeţe: prima suprafaţă cuprinsă între curbă, axa absciselor şi dreapta *a'c'*, cea de-a doua, între curbă, axa absciselor şi dreapta *ac*.

Se obişnuieşte să se noteze una din aceste suprafeţe, cea mărginită de dreapta *ac*, de pildă, prin simbolul:

$$\frac{N}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-M)^2}{2\sigma^2}} dx \qquad (43)$$

unde x este valoarea variabilei independente corespunzătoare punctului *a*, iar $-\infty$ este simbolul valorii celei mai mici cu putinţă. (Curba normală se întinde de o parte şi de alta a valorii centrale fără ca să atingă axa absciselor, aşa că suprafaţa considerată trebuie socotită ca începând de la $-\infty$ *minus infinit.*) Expresia (43) reprezintă frecvenţa teoretică cumulată până la valoarea x.

Vom avea, aşadar:

$$Supr.\ (aa'cc') = N\left[\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x+d} e^{-\frac{(x-M)^2}{2\sigma^2}} dx - \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-M)^2}{2\sigma^2}} dx\right]$$

S-au calculat valorile *integralei:*

$$S(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(x-M)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{X} e^{-\frac{X^2}{2}} dX \ ^{1} \qquad (44)$$

[1] *Valoarea lui S are aceeaşi semnificaţie ca şi (43), însă în raport cu distribuţia compusă din frecvenţele relative reduse la unitate, N = 1.*

Scanned with OKEN Scanner

pentru diferite valori ale abaterii reduse X; le găsim în Tabla II la sfârșitul cărții. Astfel, frecvența cazurilor cuprinse între valorile x_1 și x_2, reprezentată prin suprafața *aca'c'*, va fi dată de diferența:

$$\Delta S = S(X_2) - S(X_1). \tag{45}$$

Este nevoie să accentuăm faptul că pentru calculul lui S trebuie să utilizăm abaterile în raport cu media aritmetică ale limitelor intervalelor, și nu ale mijloacelor acestora.

Aceste abateri sunt cele cuprinse în coloana (2) și (3) (tabelul 67).

TABELUL 67

Ajustare graduată de tip normal pentru cazul unei distribuții de tije de orz în raport cu înălțimea

Înălțimea în cm	Abateri $x - M$	Abateri relative $(x - M)/\sigma$	$S(X)$	Diferența ΔS	Frecvențe teoretice	Frecvențe statistice
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sub 55				0,0033051	0,6	0
55-60	-35,824	-2,7259	0,0033051	0,0061990	1,1	0
60-65	-30,824	-2,3454	0,0095041	0,0152045	2,8	4
65-70	-25,824	-1,9650	0,0247086	0,0318336	5,8	8
70-75	-20,824	-1,5845	0,0565422	0,0577361	10,5	11
75-80	-15,824	-1,2041	0,1142783	0,0936615	17,0	17
80-85	-10,824	-0,8236	0,2079398	0,1184161	21,5	20
85-90	- 5,824	-0,4432	0,3263559	0,1486470	27,1	25
90-95	- 0,824	-0,0627	0,4750029	0,1496777	27,2	28
95-100	+ 4,176	+0,3178	0,6246806	0,1327917	24,2	24
100-105	+ 9,176	+0,6982	0,7574723	0,1021656	18,6	18
105-110	+14,176	+1,0787	0,8596379	0,0680924	12,4	14
110-115	+19,176	+1,4591	0,9277303	0,0393560	7,2	6
115-120	+24,176	+1,8396	0,9670863	0,0197043	3,6	5
120-125	+29,176	+2,2200	0,9867906	0,0085536	1,6	2
125-130	+34,176	+2,6004	0,9953442	0,0032187	0,6	0
Peste 130	+39,176	+2,9809	0,9985629	0,0014371	0,2	0
					182,0	**182**

$m_1 = -0{,}3352$ $m_2 = 7{,}10439560$ $m_3 = -6{,}730769231$ $m_4 = 128{,}68681319$ [1]

$M = 90{,}8240$ $\sigma = 13{,}1422$ $\beta_1 = 0{,}000347$ $\beta_2 = 2{,}533448$.

[1] *Momentele sunt măsurate în unități-intervale.*

Scanned with OKEN Scanner

În ceea ce privește calculul lui S cu ajutorul tablelor, trebuie să facem următoarea observație. Tablele II ne dau valorile suprafeței S_1 (graficul 37) cuprinse între ordonata abaterii relative X și ordonata mediei, iar nouă ne trebuie valoarea suprafeței S. Cum $S + S_1 = 0,5$ (semisuma frecvențelor relative reduse la unitate) vom obține valorile lui S dacă vom face diferența între valorile S_1, date de table, și 0,5, adică:

$$S(X) = 0,5 - S_1(X).$$

Pentru abaterile pozitive vom avea:

$$S(X) = 0,5 + S_1(X)$$

după cum este ușor să ne încredințăm, considerând pe grafic o abatere pozitivă.

Astfel, pentru abaterile relative - 0,35, +2,01 vom avea respectiv valorile lui S: 0,5 - 0,13683 = 0,36317; 0,5 + 0,47778 = 0,97778.

În cazul valorilor care nu se găsesc în table, vom face interpolare liniară, așa cum am făcut și pentru valorile lui z.

Valorile lui S astfel obținute reprezintă frecvențele cumulate corespunzătoare frecvențelor teoretice relative reduse la unitate. Aflarea acestora din urmă se bazează pe o proprietate a frecvențelor cumulate, despre care am vorbit la timp (§ 50). Frecvențele teoretice relative se obțin făcând diferențele valorilor succesive ale lui S (45) (coloana 5). În fine, prin înmulțirea acestora cu N (= 182), avem frecvențele teoretice corespunzătoare seriei noastre, calculate sub *premisa că aceasta este o distribuție normală* (coloana 6).

Acest mod de ajustare a curbei de frecvență teoretică poartă numele de *ajustare graduată*, spre deosebire de cea din paragraful precedent, care este o *ajustare negraduată.*

81. Pentru trasarea curbei frecvențelor corespunzătoare, avem nevoie să cunoaștem nu mărimea suprafețelor ce reprezintă frecvențele teoretice, ci mărimea ordonatelor curbei, corespunzătoare mijloacelor intervalelor, adică valorile lui:

$$\frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(x-M)^2}{2\sigma^2}}$$

pentru abaterile relative ale acestor mijloace în raport cu media aritmetică.

Problema prezintă o perfectă analogie cu cea a ajustării negraduate, cu simpla deosebire că nu vom mai înmulți valorile lui z cu $\frac{Nd}{\sigma}$, ci cu $\frac{N}{\sigma}$. Aceasta, din cauză că factorul d, baza dreptunghiurilor ce reprezintă frecvențele, trebuie suprimat atunci când căutăm valorile înălțimilor acestora.

Calculul amănunțit în legătură cu această problemă este arătat în tabelul 68. Prin înmulțirea valorilor ordonatelor cu d (= 5), obținem frecvențele ideale calculate printr-o ajustare negraduată.

Scanned with OKEN Scanner

Din consideraţiile de mai sus se desprinde, în primul rând, diferenţa fundamentală între ajustarea graduată şi cea negraduată. Pe când ajustarea graduată ţine seamă de faptul că frecvenţele ideale sunt reprezentate de o suprafaţă cuprinsă între două ordonate, axa absciselor şi o linie curbă, ajustarea negraduată asimilează această suprafaţă cu un dreptunghi de înălţime egală cu ordonata curbei corespunzătoare mijlocului intervalului. Bineînţeles, ajustarea graduată este cea care interpretează mai exact realitatea. Cu toate acestea, după cum se poate constata din comparaţia tabelelor 67 şi 68, nu există o diferenţă prea mare între rezultatele celor două feluri de ajustări. De aceea, din cauza calculelor mai simple pe care le comportă, ajustarea negraduată este de recomandat, afară de cazurile în legătură cu chestiuni de actuariat, unde cea mai mică diferenţă poate cauza pierderi apreciabile societăţilor de asigurare.

TABELUL 68

Ajustarea negraduată după tip normal

Intervale	Abateri $x - M$	$x - \frac{M}{\sigma}$	z	Ordonate $z \times \frac{N}{\sigma}$	Frecvenţe $z \times \frac{N}{\sigma} \times d$
(1)	(2)	(3)	(4)	(5)	(6)
50-55	-38,324	-2,9161	0,0056808	0,078	0,4
55-60	-33,324	-2,5356	0,0160263	0,222	1,1
60-65	-28,324	-2,1552	0,0391116	0,542	2,7
65-70	-23,324	-1,7747	0,0826045	1,144	5,7
70-75	-18,324	-1,3943	0,1509197	2,090	10,5
75-80	-13,324	-1,0138	0,2386319	3,305	16,6
80-85	- 8,324	-0,6334	0,3264290	4,521	22,6
85-90	- 3,324	-0,2529	0,3856834	5,341	26,7
90-95	+ 1,676	+0,1275	0,3957091	5,480	27,4
95-100	+ 6,676	+0,5080	0,3506466	4,856	24,3
100-105	+11,676	+0,8884	0,2689070	3,724	18,6
105-110	+16,676	+1,2689	0,1783533	2,470	12,4
110-115	+21,676	+1,6493	0,1023836	1,418	7,1
115-120	+26,676	+2,0298	0,0508447	0,704	3,6
120-125	+31,676	+2,4103	0,0218467	0,303	1,5
125-130	+36,676	+2,7907	0,0081241	0,113	0,6
130-135	+41,676	+3,1711	0,0026141	0,036	0,2
					182,0

Tabelul 67

82. *Abaterea-tip şi legea normală.* Distribuţia normală fiindu-ne acum întrucâtva mai familiară, vom trece la analiza câtorva proprietăţi esenţiale ale ei.

Scanned with OKEN Scanner

Examinând Tablele II, observăm, în primul rând, că suprafaţa cuprinsă între curba normală, axa absciselor, în afara intervalului egal cu 3σ de o parte şi de alta a valorii medii, este egală cu $2 \times N \times 0{,}0013499$. Ceea ce înseamnă că într-o distribuţie normală abaterile care întrec de trei ori abaterea-tip în valoare absolută nu se prezintă decât aproximativ de 26 de ori în 10.000 de observaţii. *Frecvenţa abaterilor relative mai mari de 3 în valoare absolută fiind aşa de mică, putem spune că apariţia unei astfel de valori este o cvasiimposibilitate sau, cu alte cuvinte, avem o cvasicertitudine că, într-un sondaj făcut într-o populaţie distribuită normal, toate valorile vor fi cuprinse între* $M - 3\sigma$ *şi* $M + 3\sigma$.

Aceasta este una din caracteristicile cele mai importante ale distribuţiei normale. Forma pe care o îmbracă proprietatea enunţată este o justificare în plus a importanţei teoretice şi practice a instrumentelor de cercetare care sunt media aritmetică şi abaterea-tip.

Pentru verificarea acestui fapt, să considerăm distribuţia din tabelul 67. Avem în acest caz:

$$M - 3\sigma = 51{,}397 \qquad M + 3\sigma = 130{,}251.$$

Nici o observaţie nu iese din acest interval, valorile distribuţiei considerate fiind toate cuprinse între 60 şi 125.

Proprietatea aceasta se verifică şi pentru alte distribuţii, cum sunt distribuţiile simetrice cu un coeficient de boltire care să nu difere prea mult de valoarea 3, precum şi pentru distribuţiile uşor asimetrice. Comitem însă o eroare însemnată dacă aplicăm această proprietate altor tipuri de distribuţii, cum ar fi, de pildă, cele în J sau în U. Legea de mai sus ne spune, într-adevăr, că în cazul distribuţiilor simetrice sau uşor asimetrice marea majoritate a observaţiilor se masează împrejurul mediei aritmetice a seriei. Acesta este faptul crud, forma sub care este enunţat mai sus nu vine decât să precizeze limitele între care se face această concentrare de valori. Or, însăşi forma distribuţiilor în J şi în U ne arată singură că nu poate fi vorba despre o astfel de lege în legătură cu aceste tipuri de distribuţie. Concentrarea observaţiilor se face în acest caz către una sau către ambele extremităţi ale scării distribuţiilor.

83. *Caracterizarea normalităţii prin frecvenţele intervalelor tip*. Se poate face chiar o analiză mai aprofundată a felului în care se face gruparea valorilor observate împrejurul mediei aritmetice. Din Tablele II se deduce următoarea distribuţie de frecvenţe relative, în raport cu intervale egale cu $\frac{1}{2}\sigma$ [1]:

[1] *Frecvenţele teoretice absolute se obţin prin înmulţirea acestora cu totalul frecvenţelor N.*

Scanned with OKEN Scanner

	sub $M - 3\sigma$	0,0013
între $M - 3\sigma$	şi $M - \frac{5}{2}\sigma$	0,0049
între $M - \frac{5}{2}\sigma$	şi $M - 2\sigma$	0,0166
între $M - 2\sigma$	şi $M - \frac{3}{2}\sigma$	0,0440
între $M - \frac{3}{2}\sigma$	şi $M - \sigma$	0,0919
între $M - \sigma$	şi $M - \frac{1}{2}\sigma$	0,1498
între $M - \frac{1}{2}\sigma$	şi M	0,1915
între M	şi $M + \frac{1}{2}\sigma$	0,1915
între $M + \frac{1}{2}\sigma$	şi $M + \sigma$	0,1498
între $M + \sigma$	şi $M + \frac{3}{2}\sigma$	0,0919
între $M + \frac{3}{2}\sigma$	şi $M + 2\sigma$	0,0440
între $M + 2\sigma$	şi $M + \frac{5}{2}\sigma$	0,0166
între $M + \frac{5}{2}\sigma$	şi $M + 3\sigma$	0,0049
	peste $M + 3\sigma$	0,0013
		1,0000

Vom trece acum la o comparaţie pe aceste baze între frecvenţele observate şi cele teoretice, în cazul distribuţiei din tabelul 67. Calculul necesitând în primul rând schimbarea limitelor intervalelor, vom presupune că în fiecare interval al distribuţiei primitive elementele respective se distribuie în chip uniform.

Tabelul 69 ne arată procedeul de calcul, ale cărui amănunte nu le dăm, din cauza lungimii lor.

Scanned with OKEN Scanner

TABELUL 69

Verificarea normalităţii unei distribuţii prin intervalele tip

Limitele intervalelor	Frecvenţe teoretice	Frecvenţe statistice
51,3974	0,24	0
57,9685	0,89	0
64,5396	3,02	3,63
71,1107	8,01	10,81 = 0,37 + 8 + 2,44
77,6818	16,73	17,68 = 8,56 + 9,12
84,2529	27,26	24,89 = 7,88 + 17,01
90,8240	34,85	32,60 = 2,99 + 25 + 4,61
97,3951	34,85	34,89 = 23,39 + 11,50
103,9662	27,26	26,78 = 12,50 + 14,28
110,5373	16,73	18,36 = 3,72 + 14 + 0,64
117,1084	8,01	7,47 = 5,36 + 2,11
123,6795	3,02	4,36 = 2,89 + 1,47
130,2506	0,89	0,53
	0,24	0

Tabelul 67

Ca o simplă îndrumare, vom arăta cum s-a obţinut frecvenţa statistică corespunzătoare intervalului 103,9662 - 110,5373. Acest interval se descompune în următoarele:

103,9662 - 105
105 - 110
110 - 110,5373

Frecvenţa primului va fi dedusă din frecvenţa intervalului 100-105, prin interpolare liniară:

$$(5 - 3{,}9662) \times \frac{18}{5} = 3{,}72.$$

La fel, frecvenţa ultimului interval va fi:

$$0{,}5373 \times \frac{6}{5} = 0{,}64,$$

de unde se obţine frecvenţa căutată.

84. *Controlul normalităţii coeficientului de masculinitate.* Prof. LUCIEN MARCH[1] de la Institutul de Statistică al Universităţii din Paris are o predilecţie specială pentru întrebuinţarea acestui din urmă procedeu de verificare a normalităţii, cu singura deosebire că înţelege să aprecieze astfel regularitatea unui fenomen.

TABELUL 70

Coeficientul lunar de masculinitate pentru România

Luna	1904	1905	1906	1907	1908
Ianuarie	0,5228	0,5100	0,5181	0,5174	0,5213
Februarie	0,5191	0,5160	0,5115	0,5170	0,5120
Martie	0,5161	0,5236	0,5152	0,5117	0,5217
Aprilie	0,5173	0,5160	0,5200	0,5131	0,5129
Mai	0,5181	0,5140	0,5124	0,5158	0,5159
Iunie	0,5204	0,5173	0,5193	0,5138	0,5120
Iulie	0,5164	0,5114	0,5128	0,5147	0,5104
August	0,5165	0,5111	0,5133	0,5120	0,5122
Septembrie	0,5214	0,5154	0,5171	0,5080	0,5237
Octombrie	0,5135	0,5142	0,5141	0,5116	0,5170
Noiembrie	0,5056	0,5137	0,5100	0,5108	0,5123
Decembrie	0,5287	0,5129	0,5147	0,5071	0,5046
M = 0,5151				**σ = 0,0044**	

Tabel calculat cu ajutorul datelor din ***Mişcarea populaţiei*** *pentru anii respectivi.*

March aplică de obicei acest mijloc de cercetare seriilor formate din valori individuale. Ar putea părea absurd să vorbim despre normalitatea unui şir de valori, noţiunea de normalitate fiind strâns legată de forma unei curbe de frecvenţă. Contrazicerea este numai aparentă, căci orice distribuţie heterogradă este obţinută dintr-o înregistrare statistică ce se prezintă la început sub forma unei serii de valori individuale.

Să trecem la o exemplificare prin calcul şi, în acest scop, să considerăm seria formată din coeficientul de masculinitate, calculat cu ajutorul datelor lunare din perioada 1904-1908 pentru naşterile din România[2].

Distribuţia coeficientului de masculinitate în raport cu intervalele tip este cea din tabelul 71.

---

[1] *În cursul tipăririi cărţii de faţă am primit trista veste a morţii lui Lucien March. Ştiinţa pierde în el o minte clară, un spirit critic şi un adânc cunoscător al posibilităţilor de aplicare a metodei statistice în domeniul economiei politice.*

[2] *Coeficientul de masculinitate este raportul dintre numărul de naşteri masculine şi numărul total al naşterilor.*

Scanned with OKEN Scanner

TABELUL 71

Verificarea normalităţii coeficientului de masculinitate[1]

Limitele intervalelor	Frecvenţe statistice	Frecvenţe teoretice
5019	0	0,1
5041	0	0,3
5063	2	1,0
5085	2	2,6
5107	3	5,5
5129	14	9,0
5151	11	11,5
5173	13	11,5
5195	6	9,0
5217	4,5	5,5
5239	3,5	2,6
5261	0	1,0
5283	0	0,3
	1	0,1

Tabelul 70

85. *Alte proprietăţi ale distribuţiei normale. Variaţia statistică.* Controlul normalităţii se mai poate face şi după alte criterii. Acestea sunt bazate pe unele proprietăţi ale distribuţiei normale, care constituie un indiciu de normalitate, fără a fi însă caracteristice în acest sens. În această ordine de idei, vom menţiona următoarele două egalităţi:

$$\frac{e_M}{\sigma} = 0{,}7978 \qquad (0{,}8)$$

$$\frac{Q}{\sigma} = 0{,}6745 \qquad \left(\frac{3}{2}\right) \qquad (46)$$

care, bineînţeles, nu vor fi verificate de către distribuţiile statistice decât în mod aproximativ.

Astfel, vom avea pentru distribuţia heterogradă din tabelul 66:

[1] *Când una din valori a fost egală cu limita dintre două intervale, s-a repartizat ambelor intervale frecvenţa 0,5.*

Scanned with OKEN Scanner

$$\frac{e_M}{\sigma} = \frac{1{,}1625}{1{,}4889} = 0{,}781$$
$$\frac{Q}{\sigma} = \frac{1}{1{,}4889} = 0{,}672\,, \qquad (47)$$

iar pentru cea din tabelul 67:

$$\frac{e_M}{\sigma} = \frac{10{,}7424}{13{,}1422} = 0{,}817$$
$$\frac{Q}{\sigma} = \frac{9{,}2604}{13{,}1422} = 0{,}705\,. \qquad (48)$$

În cazul distribuţiilor normale, valorile e_M, Q, σ stând într-un raport constant una faţă de alta, va fi, prin urmare, indiferent care din aceste expresii va fi întrebuinţată pentru a ne da seama de dispersia unei atare distribuţii. Alegerea a fost făcută demult sub influenţa unor rezultate din teoria probabilităţilor. Uzul a consacrat această alegere care a purtat asupra lui Q, dispersia interquartilă. Singurul motiv care a condus la această alegere a fost faptul că abaterile mai mari decât Q sunt egale în număr cu cele inferioare acestuia. Cu alte cuvinte, avem şanse egale ca să întâlnim o abatere mai mare sau mai mică decât Q. Vedem dar că dispersia normală apare sub o formă foarte intuitivă dacă se face apel la valoarea lui Q. Această valoare este notată în literatura engleză prin *P. E.*, iniţialele cuvintelor *probable error*, termenul întrebuinţat fiind în legătură cu proprietatea menţionată precedent[1].

Tabelul 72 ne arată în mod mai amănunţit cum se repartizează abaterile în raport cu o scară de multipli de Q, în cazul unei distribuţii normale.

TABELUL 72

Frecvenţele relative ale abaterilor în raport cu multiplii de Q

Abatere	Frecvenţele abaterilor ce o depăşesc
Q	0,5
$2\,Q$	0,177
$3\,Q$	0,0430
$4\,Q$	0,00698
$5\,Q$	0,000746
$6\,Q$	0,000052

[1] *Termenul franţuzesc este* écart médian. *Şi acesta reflectă aceeaşi proprietate caracteristică a lui Q.*

Scanned with OKEN Scanner

Din 20 de abateri numai una singură va depăşi în general valoarea 3Q. Raritatea cu care se produce o abatere mai mare decât 3Q a făcut ca astfel de abateri să fie considerate ca *semnificative*, nu însă ca imposibile.

Făcând o legătură între consideraţiile din acest paragraf şi § 93, ţinând seamă şi de faptul că avem aproximativ $3Q = 2\sigma$, vom rezuma un rezultat destul de important:

În cazul unei distribuţii normale, abaterile ce nu întrec 2σ sunt abateri *obişnuite*, cele cuprinse între 2σ şi 3σ sunt *semnificative*, iar cele mai mari decât 3σ, cvasiimposibile[1]. Din această cauză intervalul $M - 2\sigma$, $M + 2\sigma$ poate fi numit *interval de variaţie statistică* a fenomenelor colective.

86. *Ajustarea unei distribuţii de tipul II Pearson.* Vom examina acum un alt tip de distribuţie simetrică, cu intenţia de a oferi mai mult un exemplu numai din mulţimea de metode de ajustare care s-au imaginat. Vom avea astfel ocazia să ne dăm seama de întinderea pe care o pot lua în acest caz calculele necesare, deşi exemplul pe care îl vom examina este unul din tipurile de distribuţie a cărui ajustare trece printre cele mai simple.

Este vorba despre distribuţia simetrică platykurtică (§ 75). Tipul de curbă ce se întrebuinţează în chip general pentru ajustarea unei atare distribuţii are ca ecuaţie:

$$y = y_0\left(1 - \frac{x^2}{a^2}\right)^m \tag{49}$$

unde y_0, m, a sunt nişte constante pe care le putem calcula în funcţie de momentele distribuţiei. Originea variabilei independente este considerată egală cu media aritmetică M, aşa că x reprezintă abaterile în raport cu această valoare.

Vom considera ca suport al metodei de ajustare aceeaşi distribuţie care ne-a servit la ajustarea graduată a tipului normal. Utilizând momentele deja calculate cu acel prilej, formulele:

$$m = \frac{5\beta_2 - 9}{2(3 - \beta_2)} \qquad a^2 = \frac{2\mu_2\beta_2}{3 - \beta_2}\ ^2 \tag{50}$$

ne dau valorile:

$$m = 3{,}93015 \qquad a = 43{,}3101.$$

În ceea ce priveşte y_0, valoarea ordonatei maxime, corespunzătoare valorii medii, ea este dată de formula:

---

[1] *Este vorba aici de valoarea absolută a abaterilor.*

[2] *Demonstraţia formulelor (50), necesitând cunoştinţe mai înaintate de matematici, nu o expunem aici (§ 213).*

Scanned with OKEN Scanner

$$y_0 = \frac{N\Gamma(2m+2)}{a \cdot 2^{2m+1}\{\Gamma(m+1)\}^2}, \tag{51}$$

unde $\Gamma(x)$ este o funcţie cu o structură mai specială. Ea este cunoscută în matematici sub numele de *funcţia gamma*. Există table care să ne dea valorile acestei funcţii corespunzătoare diferitelor valori ale variabilei independente[1].

Pentru valorile mai mari ale lui x, obţinem o bună aproximaţie a valorilor lui $\Gamma(x)$, utilizând formula lui STIRLING.

$$\log\Gamma(x+1) = \log\sqrt{2\pi} + \left(x+\frac{1}{2}\right)\log x - \left(x - \frac{1}{12x}\right)\log e \tag{52}$$

care ne dă:

$$\log y_0 = \log N + (2m+1{,}5)\log(2m+1) - \left[2m+1-\frac{1}{12(2m+1)}\right]\log e - \log a -$$

$$-(2m+1)\log 2 - \log\sqrt{2\pi} - (2m+1)\log m + 2\left(m - \frac{1}{12m}\right)\log e = 0{,}71030\ ^{2}.$$

Deci ordonata maximă a curbei frecvenţelor va fi:

$$y_0 = 5{,}132.$$

O dată aceste valori obţinute, se trece la calculul valorilor ordonatelor corespunzătoare mijloacelor intervalelor distribuţiei. Procedeul de calcul este înfăţişat în tabelul 73.

Înmulţind apoi valorile ordonatelor (coloana 8) cu mărimea intervalelor, $d = 5$, obţinem frecvenţele ideale (coloana 9). Ajustarea este însă în acest caz negraduată.

Graficul 38 ne pune în evidenţă diferenţa între ajustarea după tip normal şi cea după tipul dat de ecuaţia (49), care este cunoscut sub numele de *tipul II Pearson.*

Profesorul PEARSON a fost cel dintâi care a construit o teorie omogenă a tuturor tipurilor de distribuţii ce se pot întâlni[3].

---

[1] *K. Pearson,* ***Tables for Statisticians and Biometricians****, Part I, 3rd edition, p. 58, precum şi vol. IV, VIII, IX din* ***Tracts for Computers****, Cambridge University Press.*

[2] *Unele valori necesare acestui calcul se găsesc în* Tabla III *de la sfârşitul cărţii.*

[3] *A se vedea în această privinţă* Apendice, *§ 213.*

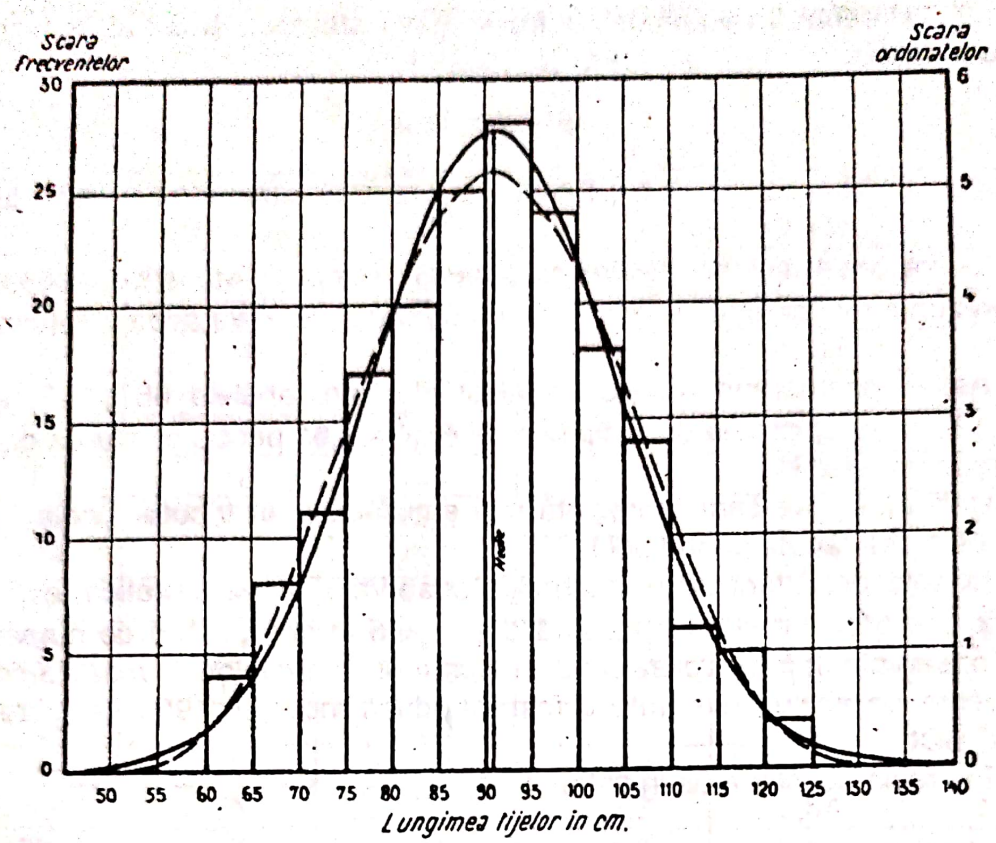
TABELUL 73

Ajustarea negraduată a tipului II Pearson

Abateri x	$\frac{x}{a}$	$\log(1-\frac{x}{a})$	$\log(1+\frac{x}{a})$	$\log(1-\frac{x^2}{a^2})$	$m\log(1-\frac{x^2}{a^2})$	$m\log(1-\frac{x^2}{a^2})+\log y_0$	Ordonate y	Frecvențe teoretice $y \times d$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
-38,324	-0,8849	0,27529	$\bar{1}$,06108	$\bar{1}$,33637	-2,60817	$\bar{2}$,10213	0,01265	0,1
-33,324	-0,7694	0,24783	$\bar{1}$,36286	$\bar{1}$,61059	-1,53005	$\bar{1}$,18025	0,1514	0,7
-28,324	-0,6540	0,21854	$\bar{1}$,53908	$\bar{1}$,75762	-0,95259	$\bar{1}$,75771	0,5724	2,9
-23,324	-0,5385	0,18710	$\bar{1}$,66417	$\bar{1}$,85127	-0,58453	0,12577	1,336	6,7
-18,324	-0,4231	0,15323	$\bar{1}$,76110	$\bar{1}$,91433	-0,33670	0,37360	2,364	11,8
-13,324	-0,3076	0,11648	$\bar{1}$,84036	$\bar{1}$,95684	-0,16963	0,54067	3,473	17,4
- 8,324	-0,1922	0,07635	$\bar{1}$,90730	$\bar{1}$,98365	-0,06426	0,64604	4,426	22,1
- 3,324	-0,0767	0,03210	$\bar{1}$,96534	$\bar{1}$,99744	-0,01006	0,70024	5,015	25,1
+ 1,676	+0,0387	$\bar{1}$,98286	0,01649	$\bar{1}$,99935	-0,00254	0,70776	5,102	25,5
+ 6,676	+0,1541	$\bar{1}$,92732	0,06225	$\bar{1}$,98957	-0,04098	0,66932	4,670	23,3
+11,676	+0,2696	$\bar{1}$,86356	0,10366	$\bar{1}$,96722	-0,12882	0,58148	3,815	19,1
+16,676	+0,3850	$\bar{1}$,78888	0,14145	$\bar{1}$,93033	-0,27380	0,43650	2,732	13,6
+21,676	+0,5005	$\bar{1}$,69854	0,17624	$\bar{1}$,87478	-0,49212	0,21818	1,653	8,3
+26,676	+0,6159	$\bar{1}$,58444	0,20841	$\bar{1}$,79285	-0,81412	$\bar{1}$,89618	0,7874	3,9
+31,676	+0,7314	$\bar{1}$,42911	0,23840	$\bar{1}$,66751	-1,30673	$\bar{1}$,40357	0,2533	1,3
+36,676	+0,8468	$\bar{1}$,18526	0,26642	$\bar{1}$,45168	-2,15497	$\bar{2}$,55533	0,03592	0,2

Tabelul 67

Scanned with OKEN Scanner



Tabelele 67, 68 şi 73

Graficul 38. Curba normală şi tip II Pearson ajustate pentru cazul distribuţiei unor tije de orz.

87. *Coeficientul de satisfacţie pentru ajustări.* Alegerea tipului de curbă de frecvenţă întrebuinţat pentru o ajustare fiind lăsată întrucâtva şi la o apreciere subiectivă a noastră, este necesar ca, o dată ajustarea făcută, să avem un mijloc obiectiv de a ne da seama dacă tipul ales reprezintă satisfăcător sau nu distribuţia considerată. Eventual, acest mijloc ne va permite să alegem din mai multe ajustări făcute pe cea mai bună.

Însemnând cu n_i', n_i frecvenţele observate şi calculate corespunzătoare aceluiaşi interval, este evident că diferenţa $n_i' - n_i$ constituie o măsură a depărtării dintre realitate şi calcul. Această diferenţă este însă o *măsură absolută.* Nu pot fi considerate la fel două diferenţe egale cu 2, provenind, prima din 4 - 2, iar a doua, din 1016 - 1014. De aceea, trebuie căutată o măsură relativă a diferenţei dintre frecvenţele observate şi cele calculate. S-a ajuns la a adopta pentru această măsură expresia:

$$\frac{(n_i' - n_i)^2}{n_i},$$

iar pentru diferența între ansamblul frecvențelor statistice și cel al frecvențelor calculate, suma:

$$\chi^2 = \sum \frac{(n_i' - n_i)^2}{n_i}. \qquad (53)$$

Dintre două ajustări, *făcute pe aceeași distribuție statistică*, aceea care va reprezenta mai bine realitatea va fi cea pentru care χ^2 va avea o valoare mai mică.

Astfel, comparând rezultatele ajustărilor din tabelele 68 și 73, avem: $\chi^2 = 5{,}42$ pentru ajustarea după tip normal și $\chi^2 = 3{,}61$ pentru ajustarea după tip II Pearson (tabelul 74).

Deci ajustarea care reprezintă mai satisfăcător distribuția lungimii tijelor de orz este cea făcută după tipul II.

Măsura depărtării dintre distribuția observată și cea teoretică, exprimată numeric prin χ^2, este și ea o măsură depinzând în primul rând de numărul de cazuri observate și în al doilea rând de numărul intervalelor. O măsură care să țină seamă de aceste diferențe a fost introdusă încă din 1900 de către prof. K.PEARSON[1].

Expresia acestei măsuri este:

$$P = \frac{\int_{\chi}^{\infty} e^{-\frac{\chi^2}{2}} \chi^{p-1}\, d\chi}{\int_{0}^{\infty} e^{-\frac{\chi^2}{2}} \chi^{p-1}\, d\chi} \qquad (54)$$

unde $p - 1$ reprezintă numărul intervalelor distribuției ideale.

P este probabilitatea ca, presupunând că fenomenul considerat se distribuie într-adevăr după legea adoptată, un alt sondaj statistic conținând același număr de intervale și de elemente să difere de distribuția teoretică mai mult decât rezultatul primei înregistrări statistice. Ajustarea va fi cu atât mai bună cu cât P va fi mai aproape de unitate. Vom numi această valoare cu termenul de *indice de satisfacție al ajustării* [2].

---

[1] *K. Pearson, **On the criterion that a given system of deviation from the probable in the case of a correlated system of variables is such that it can reasonable be supposed to have arisen from random sampling**. Philosophical Magasine, 1900, p. 157.*

[2] *În engleză, acest mijloc de a aprecia cât de reprezentativă este o ajustare este cunoscut sub numele de* Chi-Squared Test of Goodness of Fit. *Termenul propus în românește este un termen mult mai general care mi-a fost sugerat de profesorul G.Darmois în cursul unei discuții asupra indicilor întrebuințați în statistica matematică.*

Scanned with OKEN Scanner

TABELUL 74

Calculul lui χ^2

Intervale	Frecvenţe statistice	Normal	Tip II	Deviaţii normal	Deviaţii tip II
50-	0	0,4	0,1	0,40	0,10
55-	0	1,1	0,7	1,10	0,70
60-	4	2,7	2,9	0,63	0,42
65-	8	5,7	6,7	0,93	0,25
70-	11	10,5	11,8	0,02	0,05
75-	17	16,6	17,4	0,01	0,01
80-	20	22,6	22,1	0,30	0,20
85-	25	26,7	25,1	0,11	0,00
90-	28	27,4	25,5	0,01	0,26
95-	24	24,3	23,3	0,00	0,02
100-	18	18,6	19,1	0,02	0,06
105-	14	12,4	13,6	0,21	0,01
110-	6	7,1	8,3	0,17	0,64
115-	5	3,6	3,9	0,54	0,31
120-	2	1,5	1,3	0,17	0,38
125-	0	0,6	0,2	0,60	0,20
130-	0	0,2	0,0	0,20	0,00
				5,42	**3,61**

Tabelele 68 şi 73

Există table pentru aflarea imediată a acestui indice[1].
Găsim în aceste table pentru ajustarea normală:

$$p' = p + 1 = 17 \qquad \chi^2 = 5 \qquad P = 0{,}995754$$
$$p' = p + 1 = 17 \qquad \chi^2 = 6 \qquad P = 0{,}988095$$

şi, prin interpolare liniară,

$$p' = 17 \qquad \chi^2 = 5{,}42 \qquad P = 0{,}992537.$$

La fel, obţinem:

$$p' = 17 \qquad \chi^2 = 3{,}61 \qquad P = 0{,}999265$$

ceea ce înseamnă că, dacă într-adevăr distribuţia înălţimilor tijelor de orz este de tipul II, atunci, făcând 1000 de sondaje, fiecare constând din 182 de observaţii, aproximativ 999 din distribuţiile obţinute se vor abate mai mult de la

[1] *K. Pearson, **Tables for Statisticians and Biometricians**, Part I, 3rd edition, p. 26.*

Scanned with OKEN Scanner

distribuţia teoretică, decât distribuţia considerată mai sus. Diferenţa prea mică dintre cele două valori ale lui P nu ne permite să afirmăm categoric care este tipul de curbă cel mai apropiat acestui caz. Tot ceea ce putem spune este că şi tipul normal şi tipul II reprezintă destul de satisfăcător distribuţia statistică.

88. *Intervalele marginale şi calculul lui P.* Indicele P se sprijină pe consideraţii care presupun că frecvenţele intervalelor nu sunt prea mici, cu alte cuvinte, cazurile ce pot fi cuprinse într-unul din intervalele distribuţiei nu sunt o raritate. De aceea, criteriul P nu poate fi judicios aplicat unei distribuţii care ar prezenta acest aspect, mai ales că, din cauza formei expresiei care măsoară abaterea dintre frecvenţa statistică şi cea teoretică:

$$\frac{(n_i - n'_i)^2}{n_i}$$

frecvenţele mici pot da naştere la abateri foarte mari, care vor falsifica astfel valoarea corespunzătoare lui χ^2.

Referindu-ne astfel la ajustarea din tabelul 66, avem $\chi^2 = 119{,}95$, în timp ce numai partea aferentă ultimului interval este de 84,1. Mărimea acestei din urmă valori este în parte pricinuită de faptul că $n_i = 0{,}1$.

Neajunsul menţionat în aplicarea criteriului P se poate înlătura grupând intervalele marginale într-unul singur căruia să-i corespundă o frecvenţă mai ridicată. R. A. FISHER recomandă ca această grupare să se facă în aşa fel ca nici un interval să nu prezinte o frecvenţă teoretică mai mică decât 5[1]. Personal înclin spre o mai mare prudenţă; ar fi de dorit ca gruparea să se facă, în cazul când seria cuprinde un număr mare de elemente, până când orice interval va avea o frecvenţă relativă de ordinul 0,1. În cazul când seria cuprinde mai puţin de 500 de elemente, vom rămâne evident la avizul lui Fisher.

În cazul ajustării din tabelul 66, obţinem $P = 0{,}00000$, ceea ce înseamnă că florile de mac nu se distribuie normal în raport cu numărul stigmatelor. Să grupăm acum primele 3 şi ultimele 4 intervale la un loc. Vom obţine (tabelul 75) valorile:

$$p' = 8 \qquad \chi^2 = 13{,}99 \qquad P = 0{,}05139$$

valoare care oricum diferă de cea de mai sus, fără însă ca să fie suficient de mare ca să ne îndreptăţească să afirmăm că distribuţia de la care am pornit este normală.

Se vede însă din acest exemplu cum valoarea lui P poate fi ameliorată prin gruparea intervalelor *sensibile*, deşi prin micşorarea numărului p' ar trebui să obţinem o valoare mai mică a lui P.

---

[1] *R. A. Fisher, op. cit., p. 86.*

Scanned with OKEN Scanner

TABELUL 75

Calculul lui χ^2

Numărul stigmatelor	Frecvențe statistice	Frecvențe teoretice	$\frac{(n_i' - n_i)^2}{n_i}$
Sub 7	13	25,0	5,76
7	91	98,0	0,50
8	295	282,1	0,59
9	550	517,4	2,05
10	619	604,4	0,35
11	418	449,7	2,23
12	195	213,1	1,54
Peste 12	87	78,3	0,97
			13,99

Tabelul 66

Și acum, o observație. Din considerațiile de mai sus reiese că distribuția florilor de mac în raport cu numărul stigmatelor are prea puțini sorți de a fi normală. Totuși ajustarea curbei de tip normal s-a putut face fără nici o dificultate.

Este aici un fapt de ordin mult mai general și care merită să fie accentuat: *metodele de calcul, aplicarea formulelor etc. se pot efectua întotdeauna, dar aceasta nu ne poate îndreptăți să afirmăm că suntem pe calea cea bună a interpretării realității.* O mașină de calcul adună întotdeauna două numere, oricare ar fi ele, și ne dă rezultatul exact al operației; nu putem deduce de aici că cele două numere trebuia într-adevăr să fie adunate și nu împărțite între ele. Discernământul acesta nu-l poate face nici formula, nici metoda, el nu aparține decât puterii noastre de judecată.

89. *O lămurire.* Comparând egalitățile (46), (47) și (48), am fi tentați să afirmăm că distribuția din tabelul 66 este mai aproape de cea normală decât distribuția din tabelul 67. Aplicarea criteriului P ne conduce, dimpotrivă, la concluzii contrare. Explicația stă în faptul că aplicarea criteriilor de normalitate (46) nu sunt comparabile de la o serie la alta decât în cazul când acestea ar cuprinde același număr de elemente. Legea numerelor mari[1] ne spune că ne apropiem cu atât mai mult de distribuția teoretică cu cât facem mai multe observații. Distribuția din tabelul 66 cuprinde 2 268 de observații, pe când cea din tabelul 67, numai 182. Nu putem ști cum s-ar fi verificat criteriile (46) dacă această din urmă distribuție ar fi cuprins și ea tot 2 268 de elemente.

[1] *A se vedea capitolul următor.*

Scanned with OKEN Scanner

Capitolul VII

Legile hazardului

Comment oser parler des lois du hasard? Le hasard n'est-il pas l'antithèse de toute loi?

J. Bertrand, **Calcul des probabilités**

90. *Valori statistice şi valori teoretice*. Am văzut în capitolele precedente care sunt măsurile atributelor cele mai importante ale unei serii statistice, precum şi procedeele de calcul pentru determinarea lor. Aplicând aceste rezultate la studiul statistic al unui fenomen oarecare, pornind de la o serie de observaţii ale manifestărilor acestuia, am fi poate tentaţi să credem că valorile statistice astfel obţinute reprezintă măsurile efective ale caracteristicilor fenomenului studiat. Am comite astfel una din cele mai mari erori. Valorile calculate diferă într-adevăr de cele adevărate, chiar dacă precizia observaţiilor a fost ideală şi chiar dacă pentru calculul lor s-au întrebuinţat cele mai potrivite metode.

Diferenţa aceasta provine din faptul că valorile calculate corespund unui simplu eşantion obţinut la întâmplare, eşantion care evident diferă în marea majoritate a cazurilor de colectivitatea pe care o reprezintă.

Pentru a înţelege mai bine acest lucru, să ne fixăm ideile asupra unui exemplu. Să presupunem că este vorba despre un studiu al distribuţiei taliilor locuitorilor din România. Există două procedee de-a trata această problemă. Primul constă în măsurarea taliei tuturor locuitorilor[1], al doilea din măsurarea taliei numai unei părţi din populaţia ţării.

În primul caz, studiem colectivitatea însăşi, în al doilea, studiem indirect aceeaşi colectivitate, printr-un eşantion al acesteia. Nu mai este nevoie să spunem că primul procedeu este irealizabil şi că, prin urmare, nu vom fi niciodată în situaţia de a studia direct o colectivitate[2].

Iată de ce valoarea mediei aritmetice, a abaterii-tip etc. obţinute prin eşantion vor diferi de valorile aceloraşi caracteristici statistice ale colectivităţii.

Acelaşi fapt poate fi privit şi din alt punct de vedere. Să considerăm operaţiunea constând din aruncarea banului. Dacă banul aruncat este un

---

[1] *Bineînţeles, a acelora care au o talie stabilă, admitem de la 21 de ani până la 60 de ani.*

[2] *Nici chiar printr-un recensământ nu se poate înregistra absolut toată colectivitatea. Recensământul nu este altceva decât cel mai bun tip de eşantion.*

Scanned with OKEN Scanner

cilindru drept, plat, perfect omogen, nimic nu ne îndreptăţeşte să afirmăm că, în cazul când banul este aruncat la întâmplare, va cădea mai degrabă pe una din feţe decât pe cealaltă, ceea ce înseamnă că frecvenţa relativă a cazurilor când banul arată "marca" este în mod teoretic egală cu 1/2. Facem un număr oarecare de aruncări; nu vom obţine decât în cazuri cu totul rare o frecvenţă statistică egală cu valoarea teoretică 1/2. Frecvenţa statistică obţinută printr-un eşantion constând dintr-un număr limitat de aruncări va diferi în general de cea teoretică.

Diferenţa dintre valorile calculate prin eşantion şi cele teoretice (sau ale colectivităţii complete) este produsul intervenţiei întâmplării, a hazardului deci. Eşantionul care constituie o serie statistică este un şir de valori obţinute la întâmplare. Depărtarea eşantionului faţă de distribuţia teoretică, deşi ar apărea la prima vedere ca absolut liberă să ia orice valori, este totuşi supusă unor legi, *legilor hazardului.*

91. *Caracteristicile hazardului.* Vom trece acum la examinarea noţiunii de hazard, care, deşi este considerată de mulţi ca o noţiune comună, are nevoie, atunci când este tratată într-un cadru ştiinţific, să fie cercetată mai de aproape.

Aruncând un ban, spunem că banul va cădea la întâmplare cu una sau alta din feţe în sus. Alegând din listele de la birourile populaţiei un nume oarecare, vom determina la hazard un individ mai înalt sau mai scund decât 1,65 m. Pare deci că între prima dată a problemei şi rezultatul determinat de aceasta nu există nici o legătură. Cu toate acestea, pornind de la condiţiile iniţiale ale mişcării, o minte ideală ar putea calcula traiectoria banului, efectul diferitelor şocuri suferite în cădere şi ar prevedea astfel, printr-un determinism complet, care anume faţă a banului se va găsi la sfârşit deasupra. Calculele ar fi în acest caz din cele mai complicate. Pe deasupra, rezultatul va putea fi eronat din două motive:

1. Condiţiile iniţiale ale mişcării n-au fost perfect măsurate, cea mai mică variaţie a vitezei iniţiale, imperceptibilă cu cele mai fine instrumente, poate face ca banul să arate "capul" în loc de "marcă". Cea mai mică variaţie în rezistenţa aerului, chiar un curent produs de respiraţia experimentatorului, poate duce la infirmarea prevederii.
2. Cauzele care determină şi influenţează traiectoria banului sunt aşa de numeroase, majoritatea, aşa de puţin vizibile, încât neglijarea uneia singură este un fapt firesc. Acest lucru poate duce şi el la o contrazicere între rezultatul efectiv şi prevedere.

Fenomenele supuse hazardului prezintă, prin urmare, următoarele două caracteristici:

1. *Efectul este determinat de o mulţime de cauze de intensitate variabilă şi, în general, de minimă importanţă.*
2. *O variaţie infinitezimală a unei astfel de cauze poate conduce la o variaţie de un ordin de mărime apreciabil pentru efect.*

În cazul când legătura între cauză şi efect prezintă aceste caractere, vom spune că fenomenul se produce la întâmplare.

Scanned with OKEN Scanner

Această definiţie a hazardului are calitatea de a fi obiectivă şi imuabilă faţă de progresele tehnice. Ceea ce este astăzi privit ca hazard va fi considerat ca atare, orice descoperiri noi vor mai veni să îmbogăţească tehnica ştiinţifică. Cu alte cuvinte, experimentatorul care urmăreşte descoperirea unei legi este în stare, ţinând seamă de caracteristicile enunţate, să afirme în deplină cunoştinţă de cauză dacă fenomenul studiat se produce la întâmplare, în care caz producerea *individuală* a acestuia nu poate forma obiectul unei previziuni ştiinţifice, sau se produce după o lege rigidă, nedescoperită încă. Deşi legile mişcării planetelor n-au fost cunoscute până la Ticho-Brahe, nimănui nu i-a venit un moment ideea de a afirma că această mişcare s-ar face la întâmplare[1].

Jocul hazardului face, prin urmare, imposibilă orice previziune a producerii individuale a unui fenomen. Această imposibilitate dispare atunci când este vorba de previziuni în masă; tabelele de mortalitate nu pot servi la prezicerea morţii unui individ, ele se verifică cu o precizie surprinzătoare atunci când este vorba de decesele survenite într-o colectivitate.

Previziunea aceasta este posibilă tocmai din cauza existenţei unor legi ale hazardului. Cu studiul acestora se ocupă calculul probabilităţilor. Le vom cerceta pe cele mai importante.

92. *Frecvenţe şi hazard. Distribuţia binomială.* Să considerăm, în primul rând, manifestările hazardului asupra frecvenţelor statistice.

Problema se pune în chipul următor:

Dintr-o populaţie de 10.000 de locuitori se formează la *întâmplare*[2] un eşantion de 100. Dacă în populaţia primitivă frecvenţa absolută a taliilor între 1,65 m şi 1,70 m era 452, ce vom putea spune despre frecvenţa aceloraşi talii cuprinse în eşantion?

---

[1] *Analiza precedentă este o incidenţă necesară acestui capitol. Este însă departe de a fi o expunere completă. Cititorul doritor de a cunoaşte o critică aprofundată a noţiunii de hazard se poate referi cu mult folos la introducerile următoarelor tratate:*

J. Bertrand, ***Calcul des probabilités****, Gauthier-Villars, Paris, 1899;*
H. Poincaré, ***Calcul des probabilités****, Gauthier-Villars, Paris, 1912, precum şi la*
P.S. Laplace, ***Théorie analytique des probabilités****, Paris, Gauthier-Villars, 1921(Col. «Les maîtres de la pensée scientifique»);*
J. M. Keynes, ***A Treatise in Probability****, London, Macmillan, 1921.*

[2] *În tot acest capitol este vorba despre eşantioane obţinute* la întâmplare. *Acest fapt îşi are importanţa lui. Condiţiile care să garanteze alegerea la întâmplare a elementelor compunând eşantionul trebuie să fie examinate de aproape. Acestea sunt, în majoritatea cazurilor, înşelătoare. Astfel, dacă vrând să măsurăm talia medie a populaţiei din România, hotărâm să formăm un eşantion din persoanele al căror nume începe cu Z, am acorda o pondere prea mare elementelor de altă origine etnică decât cea română şi, în acest caz, eşantionul nu va mai putea fi privit ca un eşantion obţinut la întâmplare.*
Dacă am vrea iarăşi să măsurăm dărnicia locuitorilor din Bucureşti după distribuţia sumei dată ca obol chetelor de stradă, am comite iarăşi o eroare, prin excluderea celor care circulă în vehicule şi pe care chetatorii nu-i solicită.

Scanned with OKEN Scanner

Vom examina problema pe un caz construit. Să considerăm frecvența numerelor cu soț printre acelea arătate de 12 zaruri aruncate la întâmplare (tabelul 76).

O operație constând dintr-o aruncare va fi un eșantion de 12 elemente ale colectivității teoretice formate dintr-un număr infinit de aruncări ale unui zar "ideal".

TABELUL 76

Distribuția frecvențelor numerelor pare într-o aruncare de 12 zaruri

Frecvența numerelor pare	0	1	2	3	4	5	6	7	8	9	10	11	12
Frecvențe statistice	-	4	11	63	108	191	217	213	125	53	13	2	-
Frecvențe teoretice	0,2	2,9	16,1	53,7	120,9	193,4	225,6	193,4	120,9	53,7	16,1	2,9	0,2

Distribuția aceasta are o formă pe care n-am examinat-o încă. Ea poartă numele de *distribuție binomială*, din cauza legăturii formale pe care o are cu *binomul lui Newton.* Într-adevăr, dacă frecvențele sunt observate în N eșantioane de n elemente, iar frecvența relativă teoretică a cazului observat este p, atunci distribuția teoretică a frecvențelor statistice este dată de termenii desfășurării binomului $N(p + q)^n$, unde $q = 1 - p$. În cazul exemplului de mai sus:

$$N = 1\,000 \qquad n = 12 \qquad p = q = \frac{1}{2}$$

și deci frecvențele teoretice ale distribuției din tabelul 76 vor fi date de termenii desfășurării:

$$N(q+p)^{12} = Nq^{12} + NC_{12}^{1}q^{11}p + NC_{12}^{2}q^{10}p^2 + NC_{12}^{3}q^9p^3 + NC_{12}^{4}q^8p^4 +$$
$$+ NC_{12}^{5}q^7p^5 + NC_{12}^{6}q^6p^6 + NC_{12}^{7}q^5p^7 + NC_{12}^{8}q^4p^8 + NC_{12}^{9}q^3p^9 +$$
$$+ NC_{12}^{10}q^2p^{10} + NC_{12}^{11}qp^{11} + Np^{12} =$$
$$= 0{,}2 + 2{,}9 + 16{,}1 + 53{,}7 + 120{,}9 + 193{,}4 + 225{,}6 + 193{,}4 + 120{,}9 + 53{,}7 + 16{,}1 +$$
$$+ 2{,}9 + 0{,}2 = 1000$$

(Calculele precedente nu prezintă nici o dificultate; C_a^b este simbolul combinărilor simple și se obține, dacă nu avem la dispoziție table, prin *triunghiul lui Pascal).* Media distribuției statistice este $M' = 6{,}024$, în timp ce valoarea dedusă din distribuția teoretică este:

$$M = pn = 6. \qquad (55)$$

Abaterea-tip statistică este $\sigma' = 1{,}707$, aceeaşi caracteristică are valoarea teoretică:

$$\sigma = \sqrt{npq} = 1{,}732 \ ^{1}. \tag{56}$$

Dacă însă în distribuţia din tabelul 76 se înlocuiesc frecvenţele numerelor pare prin frecvenţele relative corespunzătoare $\left(0, \frac{1}{12}, \frac{2}{12}, \ldots, \frac{12}{12}\right)$, vom avea:

$$M_f = p = \frac{1}{2} \qquad \sigma_f = \sqrt{\frac{pq}{n}} = 0{,}144. \tag{57}$$

Rezultatele precedente sunt de cea mai mare importanţă. Vedem în primul rând că, deşi frecvenţa numerelor pare obţinute dintr-o aruncătură de 12 zaruri, considerată în mod individual, nu este supusă nici unei legi rigide, în ansamblu, această frecvenţă se distribuie după o lege bine determinată. Iată deci o primă manifestare a hazardului, exprimată sub forma unei legi. Apropierea între distribuţia statistică şi cea teoretică este cât se poate de satisfăcătoare ($P = 0{,}72$).

Să examinăm acum câteva aspecte ale distribuţiei considerate. În primul rând, observăm că, în intervalul egal cu 2σ de o parte şi de alta a mediei 6, sunt cuprinse 961,6 cazuri din 1000, iar dacă mărim intervalul de o parte şi de alta cu σ, atunci toate cazurile vor fi cuprinse în acesta din urmă (cu excepţia celor corespunzătoare frecvenţelor 0 şi 12, adică 0,5 din 1000). Constatarea aceasta este evident valabilă şi pentru frecvenţele relative.

De aici deducem o proprietate pe care am mai întâlnit-o, şi anume:

Diferenţa între frecvenţa statistică şi cea teoretică nu depăşeşte decât în mod excepţional de două ori abaterea-tip corespunzătoare şi este cvasi-imposibil ca această diferenţă să fie mai mare decât de trei ori aceeaşi abatere.

Comparând acum abaterile tip (56) şi (57), vedem că dacă n, adică numărul elementelor cuprinse în eşantion, creşte, abaterea-tip a frecvenţelor absolute creşte şi ea, în timp ce aceea a frecvenţelor relative descreşte. Ceea ce înseamnă că dacă eşantionul este mai numeros, intervalul în care poate varia frecvenţa relativă se restrânge din ce în ce. Nu tot acelaşi lucru se întâmplă cu intervalul de variaţie statistică a frecvenţei absolute. Aşadar, numai frecvenţa relativă converge către frecvenţa teoretică dacă n creşte. Printre neiniţiaţi circulă opinia falsă că acest fapt se întâmplă şi cu frecvenţa absolută.

93. *Latura statistică a variaţiei frecvenţelor.* Astfel cum am examinat problema distribuţiei frecvenţelor dintr-un eşantion, am ocolit în paragraful precedent, din cauza exemplului artificial construit, o dificultate care totuşi este

---

[1] *Expresia valorilor teoretice M şi σ se poate stabili printr-o demonstraţie matematică. Se poate face însă şi o verificare numerică directă a acestor formule.*

Scanned with OKEN Scanner

de cea mai mare importanță în practică. Este vorba despre faptul că în general valoarea teoretică *p* ne este necunoscută; de multe ori chiar, sondajul statistic n-are alt scop decât o evaluare a acesteia. Știm, de exemplu, printr-un sondaj făcut, că din 2 000 de copii nou-născuți, 931 au avut o greutate la naștere mai mică decât 3 000 g. Pentru a găsi intervalul de variație statistică a frecvenței considerate, trebuie să cunoaștem valoarea teoretică a acesteia din urmă. Or, problema se pune exact în mod invers: prin cunoașterea acestui interval vrem să determinăm valoarea teoretică a frecvenței.

Pentru învingerea acestei dificultăți, se obișnuiește să se înlocuiască formula teoretică (57) printr-alta care să ne permită o evaluare apropiată a abaterii-tip căutate:

$$\sigma' = \sqrt{\frac{f(1-f)}{n}} \tag{58}$$

unde *f* este frecvența relativă statistică. Cum aceasta nu diferă prea mult de *p*, mai ales dacă *n* este suficient de mare (§ 92), nici σ' nu va diferi prea mult de σ.

În exemplul numeric de mai sus, $f = 0{,}4655$ și deci $\sigma' = 0{,}0112$, de unde deducem că numai în caz excepțional eșantionul considerat provine dintr-o colectivitate teoretică pentru care *p* este în afara intervalului 0,4431 - 0,4879[1].

Astfel revenim la cele spuse la începutul acestui capitol: valorile deduse din eșantion nu sunt cele reale, însă, după cum vedem, ele ne pot furniza informații utile asupra valorilor teoretice corespunzătoare.

94. *Regiunea de siguranță a curbelor ajustate.* O altă aplicație a rezultatelor de mai sus se poate face în legătură cu ajustarea unui tip de frecvență. După cum am văzut (Cap. VI), operația de ajustare constă din determinarea unor frecvențe teoretice $n_1, n_2,\ldots,n_p$ corespunzătoare intervalelor a căror frecvențe statistice erau respectiv $n_1', n_2',\ldots, n_p$. Cu cât valorile $n_1, n_2,\ldots,n_p$ diferă mai puțin în ansamblul lor de $n_1', n_2',\ldots, n_p'$ cu atât ajustarea este mai demnă de încredere. Am văzut cum, cu ajutorul indicelui de satisfacție *P*, putem aprecia gradul de apropiere între distribuția statistică și cea ajustată. O indicație, mai ales de ordin negativ, o putem obține plecând de la concluziile paragrafului 92.

Să însemnăm cu *N* numărul elementelor eșantionului, $N = \Sigma n_i'$. Atunci abaterea-tip a frecvenței corespunzătoare intervalului de indice *i* va fi:

$$\sigma_i = \sqrt{N \times \frac{n_i}{N}\left(1 - \frac{n_i}{N}\right)} \tag{59}$$

[1] *Este aici o inversiune de roluri care nu trebuie să pară curioasă. Dacă diferența între* f *și* p *în valoare absolută nu poate fi în mod obișnuit mai mare decât* 2σ, *atunci, cunoscând pe f, evident că:*

$$f - 2\sigma \leq p \leq f + 2\sigma.$$

și deci, în mod obișnuit, vom avea:

$$n_i' - 2\sigma_i \leq n_i \leq n_i' + 2\sigma_i.$$

Dacă se întâmplă însă ca n_i să fie exterior intervalului $n_i' - 3\sigma_i$, $n_i' + 3\sigma_i$, avem un motiv să credem că tipul de curbă ales pentru ajustare nu este cel indicat. Nu vom aplica acest principiu intervalelor pentru care valorile frecvenței sunt de ordin mai mic decât 0,1, și aceasta din motive pe care le vom vedea mai târziu (§ 96).

În practică se procedează astfel:

Se calculează pentru fiecare interval $2\sigma_i$ (tabelul 77) și se vede apoi dacă diferența în valoare absolută între n_i și n_i' întrece sau nu această valoare. Procedeul poate fi completat printr-o reprezentare grafică (graficul 39). Pornind de la vârfurile poligonului de frecvență, se ia în sus și în jos pe ordonată intervale egale cu valoarea respectivă a lui $2\sigma_i$. Se obțin astfel o serie de puncte superioare și o serie de puncte inferioare, care unite între ele prin linii poligonale determină de o parte și de alta a poligonului de frecvență o bandă de variație statistică a curbei de frecvență ajustată din interiorul căreia aceasta din urmă nu trebuie să iasă, în caz când ajustarea este bună. Este recomandabil să se traseze, în același timp, și liniile poligonale corespunzătoare intervalelor $3\sigma_i'$.

TABELUL 77

Valorile abaterilor-tip ale frecvențelor ajustate

Frecvențele statistice n_i'	Frecvențe teoretice n_i	Diferența $\lvert n_i - n_i' \rvert$	Interval de variație $2\sigma_i$
(1)	(2)	(3)	(4)
0	0,3	0,3	1,1
1	3,0	2,0	3,5
12	21,7	9,7	9,3
91	98,0	7,0	19,3
295	282,1	12,9	31,4
550	517,4	32,6	40,0
619	604,4	14,6	42,1
418	449,7	31,7	38,0
195	213,1	18,1	27,8
54	64,3	10,3	15,8
25	12,4	12,6	7,0
5	1,5	3,5	2,4
3	0,1	2,9	2,0

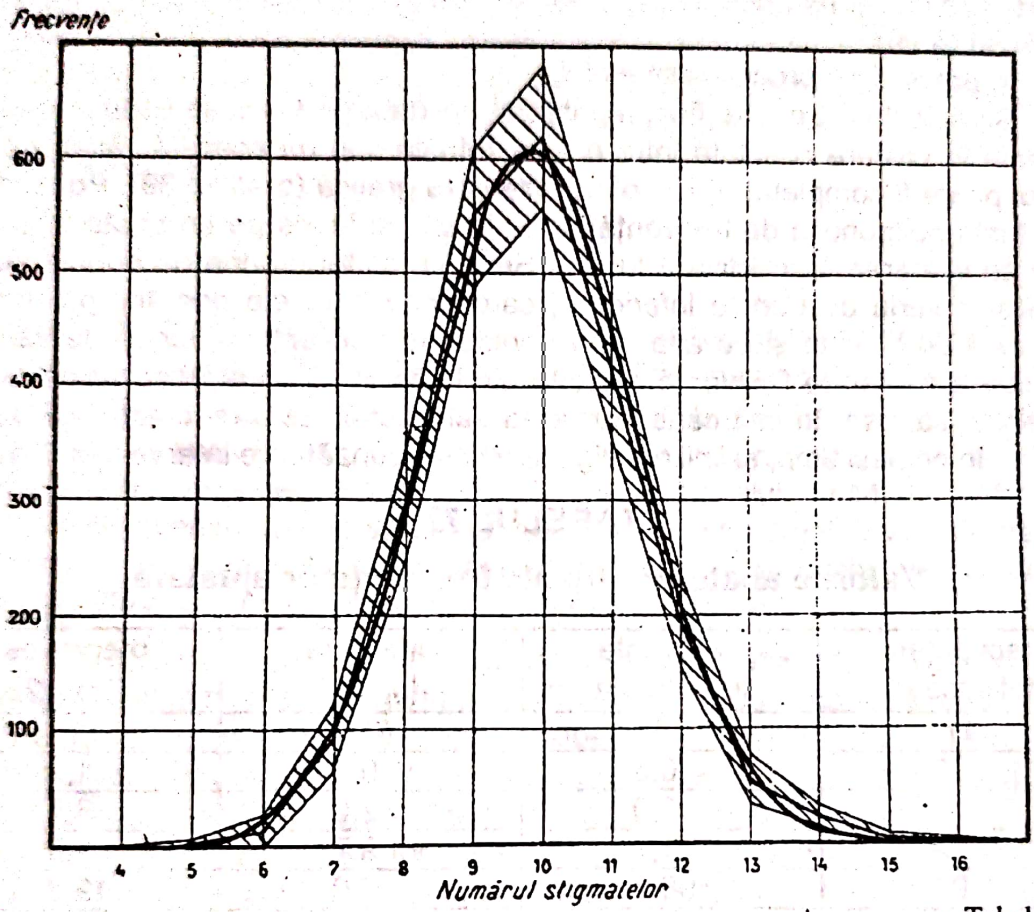
Tabelul 66

Această metodă de apreciere a ajustării este în special utilă în cazul când ajustarea se face pe cale grafică (curbă trasă cu mâna). În acest caz însă, valorile n_i ne sunt necunoscute, de aceea se înlocuiește σ_i prin:

Scanned with OKEN Scanner

$$\sigma'_i = \sqrt{N \times \frac{n'_i}{N}\left(1 - \frac{n'_i}{N}\right)}. \qquad (59 \text{ bis})$$

Pentru ordinul de precizie obținut prin ajustarea grafică, σ' înlocuiește cu succes valoarea mai judicioasă a lui σ.



Partea haşurată reprezintă regiunea de variaţie statistică.

Tabelul 77

Graficul 39. Regiunea de variație statistică a unei curbe ajustate.

95. *Distribuția binomială și legea normală*. Dacă am reprezenta grafic distribuția teoretică din tabelul 76, am observa cât de mare asemănare prezintă aceasta față de distribuția normală. Această apropiere între distribuția binomială și cea normală nu este o simplă impresie vizuală, ci o realitate care are motive mai adânci ce se pot evidenția prin procedee matematice. Ne vom mulțumi deocamdată cu o simplă verificare a acestui fapt.

În primul rând, vom înlocui distribuția în raport cu valorile individuale printr-alta în raport cu intervale egale, astfel ca fiecare valoare să devină centrul unui interval nou. Distribuția astfel prezentată se poate supune unei operații de ajustare normală graduată (tabelul 78).

Scanned with OKEN Scanner

TABELUL 78

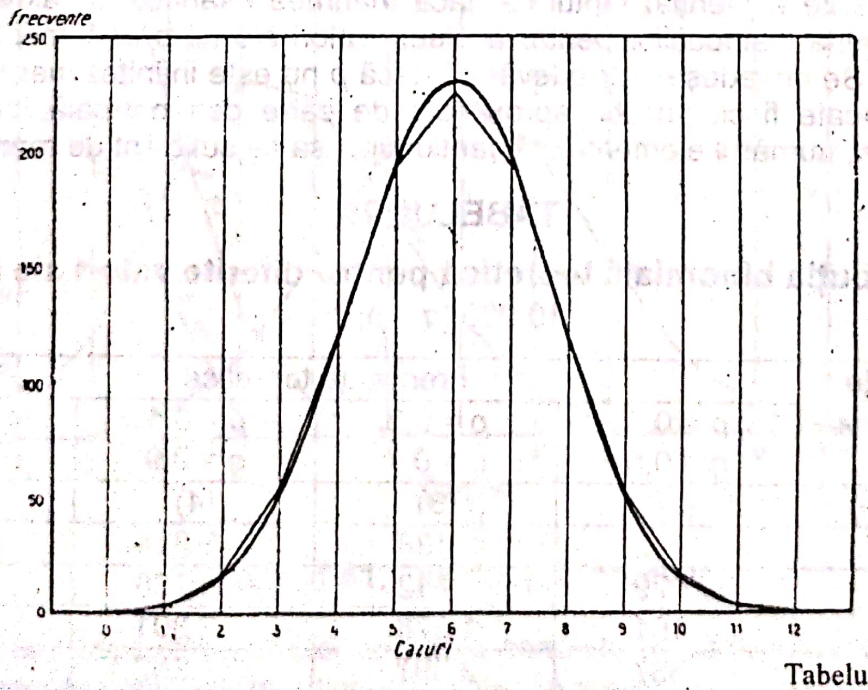
Ajustarea normală graduată pe tipul binomial

Intervale	Ordonate	Frecvențe normale	Frecvențe binomiale
(1)	(2)	(3)	(4)
-0,5-0,5	0,6	0,7	0,2
0,5-1,5	3,6	3,7	2,9
1,5-2,5	16,0	16,7	16,1
2,5-3,5	51,4	53,3	53,7
3,5-4,5	118,3	118,8	120,9
4,5-5,5	195,0	193,2	193,4
5,5-6,5	230,3	227,2	225,6
6,5-7,5	195,0	193,2	193,4
7,5-8,5	118,3	118,8	120,9
8,5-9,5	51,4	53,4	53,7
9,5-10,5	16,0	16,7	16,1
10,5-11,5	3,6	3,7	2,9
11,5-12,5	0,6	0,7	0,2

Tabelul 76

$M = 6$ $\sigma = 1{,}732$ $N = 1000$

Apropierea între cele două distribuții considerate se poate constata pe grafic (graficul 40).



Tabelul 78

Graficul 40. Comparație între distribuția binomială și curba normală.

Această apropiere între distribuția binomială și cea normală este cu atât mai accentuată cu cât eșantionul este mai numeros, adică n mai mare. Rămânând, așadar, în domeniul eșantioanelor bune (n suficient de mare), putem afirma, fără să comitem o eroare apreciabilă, că *frecvențele se distribuie după legea normală*.

96. *Asimetria distribuțiilor binomiale*. Concluziile de mai sus au avut ca suport distribuția de frecvențe din tabelul 76, în care frecvența teoretică era $p = \frac{1}{2}$. Rămâne să examinăm acum dacă apropierea între distribuția binomială și cea normală se poate face și în cazul când $p \neq \frac{1}{2}$.

Pentru aceasta, să examinăm distribuțiile binomiale teoretice din tabelul 79, precum și oglinda lor din graficul 41.

Prima constatare care se desprinde din acest examen este faptul că distribuțiile binomiale devin asimetrice pe măsură ce p crește sau descrește. (Pentru acest din urmă caz, nu este nevoie de vreo ilustrare particulară, deoarece când p crește, q descrește, iar în formula distribuției binomiale rolurile lui p și q pot fi intervertite).

Tot din examinarea formei distribuțiilor din graficul 41 se ajunge la concluzia destul de importantă că, în cazul când $p > \frac{1}{2}$, abaterile negative ale frecvenței statistice sunt mai probabile decât cele pozitive și viceversa.

Revenind asupra apropierii dintre curba normală și distribuția binomială, este bine să fie evidențiat faptul că dacă mărimea eșantionului ar fi fost mai mare, asimetria distribuțiilor pentru aceleași valori ale lui p ar fi fost mai puțin pronunțată. Se dovedește într-adevăr că dacă p nu este infinitezimal, distribuția binomială poate fi cu succes aproximată de către cea normală, cu singura condiție ca n, numărul elementelor eșantionului, să fie suficient de mare.

TABELUL 79

Distribuția binomială teoretică pentru diferite valori ale lui p

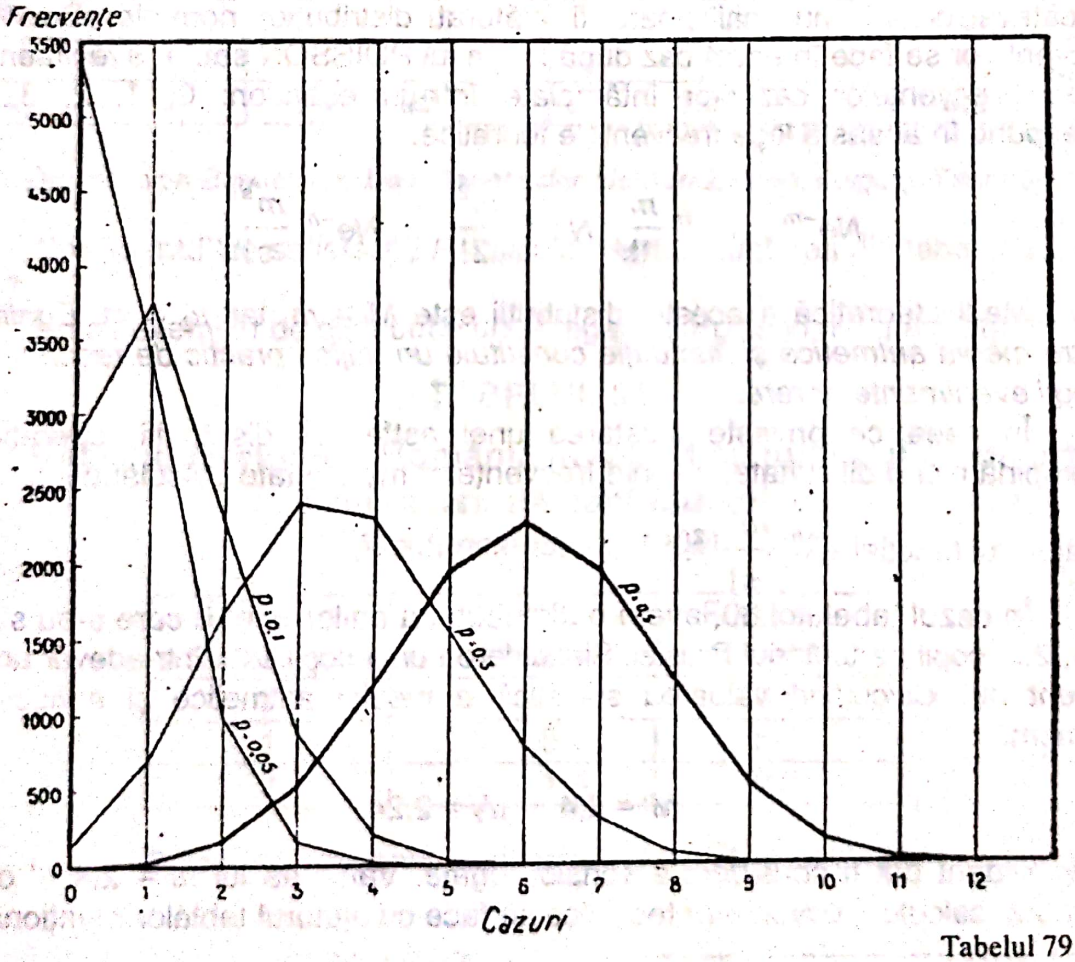
$10.000\,(q + p)^{12}$

Frecvențe în eșantion	Frecvențe teoretice			
	$p = 0,5$ $q = 0,5$	$p = 0,3$ $q = 0,7$	$p = 0,1$ $q = 0,9$	$p = 0,05$ $q = 0,95$
(1)	(2)	(3)	(4)	(5)
0	2	138	2 824	5 404
1	29	712	3 766	3 413
2	161	1 678	2 301	988
3	537	2 397	852	173
4	1 209	2 311	213	20
5	1 934	1 585	38	2

Scanned with OKEN Scanner

(continuare tab. 79)

(1)	(2)	(3)	(4)	(5)
6	2 256	793	5	-
7	1 934	291	1	-
8	1 209	78	-	-
9	537	15	-	-
10	161	2	-	-
11	29	-	-	-
12	2	-	-	-
Total	**10 000**	**10 000**	**10 000**	**10 000**
$\sqrt{\beta_1}$	0,	0,252	0,770	1,192
β_2	2,834	2,897	2,574	1,746



Tabelul 79

Graficul 41. Distribuții binomiale.

Proprietatea afirmată aici se poate constata în parte și din forma coeficienților pearsonieni ai distribuției binomiale, care sunt:

Scanned with OKEN Scanner

$$\sqrt{\beta_1} = \frac{q-p}{\sqrt{npq}} \qquad \beta_2 = 3 + \frac{1-6pq}{npq}. \tag{60}$$

O primă constatare este faptul că β_1, β_2 tind către valorile corespunzătoare distribuției normale, respectiv 0 și 3, dacă n crește, iar p (sau q) nu are o valoare prea mică. Se mai deduce din aceleași formule că apropierea distribuției binomiale față de curba normală se poate face cel mai ușor în cazul când $p = q = \frac{1}{2}$:

$$\beta_1 = 0 \qquad \beta_2 = 3 - \frac{2}{n}.$$

97. *Legea evenimentelor rare.* Cazul limită când p (sau q) au valori extrem de mici, astfel încât numărul teoretic de cazuri întâmplate np se reduce la câteva unități, nu mai poate fi alăturat distribuției normale. Distribuția frecvențelor se face în acest caz după legea lui POISSON sau a *evenimentelor rare*[1]. Frecvențelor cazurilor întâmplate într-un eșantion: 0, 1, 2, 3,... le corespund în această lege frecvențele teoretice:

$$Ne^{-m},\ Ne^{-m}\frac{m}{1!},\ Ne^{-m}\frac{m^2}{2!},\ldots,\ Ne^{-m}\frac{m^s}{s!},\ldots$$

Media teoretică a acestei distribuții este $M = m$, iar $\mu_2 = m$. *Egalitatea dintre media aritmetică și fluctuație constituie un mijloc practic de recunoaștere a legii evenimentelor rare.*

În ceea ce privește ajustarea unei astfel de distribuții, operația nu întâmpină nici o dificultate, valorile frecvențelor menționate calculându-se direct cu ajutorul funcției $e^{-m}\frac{m^s}{s!}$ [2].

În cazul tabelului 80, avem o distribuție a anilor pentru care s-au sinucis 0, 1, 2,... copii pe teritoriul Prusiei. Sinuciderea unui copil este într-adevăr un *eveniment rar.* Calculând valoarea statistică a mediei aritmetice și a fluctuației, obținem:

$$M' = 2{,}4 \qquad \mu'_2 = 2{,}24$$

care evident pot fi considerate sensibil egale. Valoarea lui $m = 2{,}4$ [3] o dată obținută, calculul frecvențelor teoretice se face cu ajutorul tablelor menționate.

---

[1] S. D. Poisson, ***Recherches sur la probabilité des jugements,*** *Paris, Bachelier, 1837.*

[2] ***Tables for Statisticians and Biometricians,*** *Part I, p. 113-121.*

[3] *Din cauză că intervalul de variație statistică a lui M' este mult mai restrâns decât cel al lui μ'_2, este mai logic să adoptăm pentru m valoarea dată de media aritmetică decât cea dată de fluctuație (§ 103).*

Scanned with OKEN Scanner

TABELUL 80

Distribuția numărului de ani în raport cu numărul copiilor sinucișI

(Prusia, 1869 - 1893)

Numărul anual al sinuciderilor	Numărul anilor	$e^{-m}\frac{m^s}{s!}$	Frecvențe teoretice $N \times (3)$
(1)	(2)	(3)	(4)
0	3	0,09	2,25
1	3	0,22	5,50
2	9	0,26	6,50
3	4	0,21	5,25
4	4	0,13	3,25
5	1	0,06	1,50
6	1	0,02	0,50
7	0	0,01	0,25
	25	**1,00**	**25,00**

*După L. von Borkiewicz, **Das Gesetz der kleinen Zahlen**, Leipzig, Teubner, 1898.*

Coeficientul de satisfacție P, calculat pentru ajustarea din tabelul 80, este 0,807.

Un alt exemplu de distribuție după legea lui Poisson ni-l oferă tabelul 81.

TABELUL 81

Distribuția județelor din România în raport cu numărul de divorțuri prin consimțământ mutual

(Comunele rurale, 1900)

Numărul divorțurilor	Numărul județelor	Frecvențe teoretice
0	26	25,6
1	5	5,8
2	1	0,6

*Datele din **Mișcarea populației României pe anul 1900**.*

$$M' = 0{,}22 \qquad \mu'_2 = 0{,}23.$$

Cu acest prilej vom observa că *raritatea* producerii unui eveniment este o noțiune relativă. După cum spuneam mai sus, un eveniment este considerat ca *rar* atunci când *np* este un număr cuprinzând câteva unități doar, ceea ce

înseamnă că nu putem face nici o afirmaţie în această direcţie numai prin cunoaşterea valorii *p*, singura care este în definitiv o caracteristică a evenimentului considerat. Raritatea unui eveniment, privită din punctul de vedere al aplicării legii lui Poisson, depinde de numărul *n* al elementelor din eşantion şi poate, prin urmare, diferi pentru acelaşi fenomen, de la un caz la altul; un fenomen (ca cel de mai sus) poate fi *rar* pentru observaţiile purtând asupra populaţiei dintr-un judeţ - şi în acest caz se distribuie după legea lui Poisson - dar poate înceta de a mai fi rar dacă observaţia se extinde la populaţia unui continent - când, prin urmare, frecvenţele se vor distribui după legea binomială sau cea normală[1].

98. *Probabilitate şi frecvenţă*. Proprietăţile distribuţiei frecvenţelor, aşa cum le-am expus în paragrafele precedente, au condus la noţiunea de *probabilitate*.

Frecvenţele obţinute în diferitele eşantioane ale aceleiaşi colectivităţi prezintă o stabilitate în modul cum se produc, stabilitate care ne îndreptăţeşte să considerăm frecvenţele statistice ca nişte evaluări ale unei mărimi ideale. Această mărime poartă numele de probabilitate şi este, ca şi frecvenţa relativă, cuprinsă între 1 şi 0. Noţiunea aceasta nu mai apare astfel prezentată ca o construcţie pur teoretică şi fără contact cu realitatea. Modul acesta de a considera probabilitatea ca izvorâtă din interpretarea unor legi ale naturii este cel care azi este admis de cei mai însemnaţi statisticieni. Este drept că, până la o anumită vreme, s-a căutat să se definească probabilitatea unui eveniment prin consideraţii de alt ordin decât cel statistic. Cu toate acestea, chiar atunci, lumea îşi dădea seama că probabilitatea este o mărime fizică şi nu un simplu coeficient numeric. Când abatele Galiani, cu toată rasa pe care o purta, a exclamat "Sangue di Bacco! Zarurile sunt măsluite", după ce un necunoscut a obţinut numai *şase* din cinci aruncări de câte 3 zaruri, înseamnă că îşi dădea destul de bine seama că probabilitatea nu este o ficţiune.

Experienţa de toate zilele, chiar dacă nu este strânsă sub forma unei serii statistice, ne învaţă că, aruncând un ban la întâmplare, vom obţine de un număr aproape egal de ori "capul" şi "marca". De aici şi până la a ne înşela pe noi înşine definind probabilitatea ca *raportul dintre numărul de cazuri favorabile şi cel al cazurilor posibile*[2] este o diferenţă apreciabilă. Fenomenul a existat

---

[1] *Raritatea unui fenomen determină pe unii statisticieni ca în publicaţiile oficiale să restrângă cadrul clasificării statistice, cu scopul de a evita înregistrarea cazurilor rare. După consideraţiile de mai sus, înclin să cred că acest procedeu nu este întotdeauna justificat. În primul rând, ceea ce este rar în România poate înceta de a mai fi la fel în Statele Unite, în al doilea, faptul că seria statistică se reprezintă sub o formă eratică este o simplă aparenţă, distribuţia frecvenţelor fiind supusă unei legi bine determinate; în al treilea rând, în fine, este un criteriu bizar de a judeca importanţa unui fenomen după frecvenţa producerii lui. Este o diferenţă între a urmări, de pildă, sinuciderile sub 14 ani într-o statistică demografică şi a înregistra într-o statistică a comerţului exterior importului acelor cu gămălie portocalie.*

[2] *Definiţie care trebuie completată prin supărătorul adaos:«dacă toate cazurile sunt la fel de probabile».*

înainte de definiția probabilității și numai experiența, prin rezultatele ei, ne-a permis să dăm o formă mai precisă consecințelor ce se desprindeau din ea.

Se obișnuiește totuși să se caute explicații pentru faptele observate în mod statistic, explicații bazate pe anumite ipoteze formulate tot în urma rezultatelor observației directe a fenomenelor colective. Se poate construi o analiză, destul de interesantă în sine, a faptului că banul aruncat prezintă cam de același număr de ori cele două fețe, însă astfel de analize nu trebuie privite decât ca niște completări explicative ale legii observate.

Evident că forma abstractă a definiției probabilității este utilă calculului probabilităților, care, ca toate ramurile științei matematice, are nevoie de entități mai epurate. Discuția precedentă are însă importanța ei, deoarece este bine să examinăm realitatea probabilității statistice înainte de a trece la orice speculație asupra probabilității teoretice.

99. *Variabilă întâmplătoare*. Vom trece acum să examinăm mai de aproape o noțiune despre care am mai vorbit și în Cap. II. Este vorba despre *variabila întâmplătoare*.

Se numește, în general, variabilă entitatea matematică prin care se exprimă o mărime oarecare. După modul în care noi putem determina efectiv valorile lor, variabilele se împart în două clase: *sigure și întâmplătoare*. Variabilele sigure reprezintă mărimile a căror intensitate o putem fixa după voință. Astfel, în limitele posibilității tehnice, noi putem obține orice temperatură vrem, putem construi un motor de forță dată etc. Nu putem însă obține *când dorim noi* un șir de 50 de aruncări de ban al căror rezultat să fie numai "capul", nu putem spune în care an viitor coeficientul de masculinitate pentru născuții din România va fi precis 0,5208, nici în ce an se vor produce exact 378529 decese pe teritoriul țării noastre. Mărimile de felul celor din urmă sunt reprezentate prin *variabile întâmplătoare*.

Variabila întâmplătoare este caracterizată prin faptul că diferitele ei valori nu se produc după voința noastră, ci una oarecare dintre ele se prezintă într-un ansamblu de experiențe de un număr de ori care este legat de valoarea individuală considerată. Să considerăm o variabilă întâmplătoare care poate lua valorile $x_1, x_2, ..., x_n$; aceste valori sunt însoțite de probabilitățile producerii lor: $p_1, p_2, ..., p_n$. Adică, dacă vom face un șir de experiențe asupra producerii fenomenului respectiv, frecvența relativă a cazurilor în care valoarea observată a fost egală cu x_1 nu va diferi cu mult de p_1 și așa mai departe. Referindu-ne la distribuția din tabelul 76, avem o variabilă întâmplătoare: numărul de numere pare obținute într-o aruncătură de 12 zaruri. Valorile pe care le poate lua această variabilă sunt: 0, 1, 2,..., 10, 11, 12. Probabilitățile producerii acestor valori sunt respectiv: 0,0002; 0,0029; 0,0161; 0,0537; 0,1209;.....; 0,0002.

În experiența făcută, frecvențele relative nu diferă sensibil de aceste valori teoretice.

100. *Problema eșantioanelor (Sampling problem)*. Exemple de variabile întâmplătoare se pot multiplica oricât de mult. Vom stărui în special asupra celor care au o importanță în statistica matematică.

Scanned with OKEN Scanner

Pentru a ne fixa ideile, să presupunem că avem de-a face cu o colectivitate compusă din 100 000 de indivizi a căror talie medie vrem s-o măsurăm printr-un eşantion de 5 000. Făcând un prim sondaj, obţinem o distribuţie heterogradă. Pornind de la această distribuţie, calculăm media aritmetică corespunzătoare:

$$M' = \sum f' x .$$

Această valoare nu va fi în general egală cu media aritmetică:

$$M = \sum px$$

a colectivităţii primitive, pentru simplul motiv că frecvenţele relative f' ale eşantionului diferă de frecvenţele p ale colectivităţii. Un alt eşantion ne va da o altă medie aritmetică:

$$M'' = \sum f'' x$$

diferită de M'.

Vedem dar că valoarea mediei aritmetice diferă de la un eşantion la altul; depinzând de frecvenţele statistice ale eşantionului, care sunt nişte variabile întâmplătoare, va prezenta şi ea simptomele hazardului. Presupunând că din aceeaşi colectivitate extragem un număr de 1000 de eşantioane, vom obţine 1000 de valori pentru media aritmetică statistică, în genere diferite între ele. Aceste 1000 de valori vor fi evaluările statistice ale unei mărimi teoretice, media aritmetică a colectivităţii sondate prin eşantioane. Valoarea mediei aritmetice a unui eşantion nu poate fi, pe deasupra, determinată *a priori* după bunul nostru plac. Aşadar, această valoare nu este altceva decât tot o variabilă întâmplătoare. Legea după care se produc diferitele valori ale mediei aritmetice statistice poate fi obţinută din formarea unei distribuţii heterograde cu ajutorul valorilor experimentale în acelaşi mod cum am procedat pentru frecvenţe (tabelul 76).

Consideraţiile de mai sus se pot aplica, mutatis mutandis, *tuturor valorilor caracteristice* ale unei distribuţii.

Prin urmare, mediana, quartilele, modul, coeficienţii de dispersie, asimetrie şi boltire, priviţi din punctul de vedere al jocului hazardului, sunt nişte variabile întâmplătoare.

Am văzut care este legea de distribuţie a variabilei întâmplătoare frecvenţă. Cunoaşterea legilor de distribuţie ale celorlalţi coeficienţi statistici constituie, şi azi încă, una din problemele cele mai interesante, dar şi cele mai greu de rezolvat, ale statisticii matematice. Problema aceasta a făcut obiectul unor cercetări asidue din partea unei bune părţi a statisticienilor şcolii engleze, dintre care, pentru a nu ţine seamă decât de lucrurile cele mai recente, vom menţiona pe J. WISHART şi R. A. FISHER.

Limbajul tehnic englez, care se adaptează aşa de uşor tuturor necesităţilor, a găsit pentru problema înfăţişată în acest paragraf termenul

Scanned with OKEN Scanner

sugestiv de *sampling problem*. Prin *sampling problem* se înţelege, aşadar, cercetarea modului în care o colectivitate poate să fie studiată printr-un eşantion al ei obţinut *la întâmplare*.

101. *Distribuţia caracteristicilor statistice. Legea eşantioanelor numeroase.* Se dovedeşte pe căi nu tocmai simple că legile de distribuţie ale valorilor caracteristice ale unui eşantion, cum ar fi media aritmetică sau mediana, de pildă, *depind de forma distribuţiei colectivităţii primitive.* Cu alte cuvinte, media aritmetică statistică va avea o lege de distribuţie în cazul când eşantionul provine dintr-o colectivitate ce se distribuie normal, şi alta, în cazul când colectivitatea primitivă va avea o distribuţie în formă de J. Această dependenţă a legii de distribuţie a unei valori statistice vine să complice şi mai mult problema determinării ei.

S-a obţinut, cu toate acestea, un rezultat general valabil pentru toate valorile statistice şi pentru orice formă de distribuţie a colectivităţii teoretice şi care este de cea mai mare importanţă pentru aplicaţiile practice ale statisticii matematice. Se dovedeşte anume că, *în cazul când eşantionul cuprinde un număr suficient de mare de elemente, aproape toate valorile coeficienţilor statistici calculaţi prin eşantion au o distribuţie normală*[1].

Forma sub care am enunţat acest rezultat fundamental are nevoie de unele interpretări. Întâlnim aici expresia "număr suficient de mare". Care este criteriul după care putem judeca dacă această condiţie este îndeplinită sau nu? Este greu de spus. În general, trebuie să convenim că un eşantion care cuprinde un număr mai mare de 500 de elemente intră în categoria celor cărora li se poate aplica legea enunţată. Este totuşi bine să fim ceva mai exigenţi în privinţa mărimii acestui număr, în cazul când colectivitatea primitivă se depărtează prea mult de forma distribuţiei normale, cum este cazul distribuţiilor pronunţat asimetrice, în J sau în U.

102. *Legea de variaţie statistică a mediei aritmetice.* Rezultatul enunţat mai sus prezintă, în afară de faptul că vedem supuse aceleiaşi legi toate valorile statistice, un alt avantaj decurgând din structura legii normale.

În primul rând, legea normală ne permite stabilirea unui interval de variaţie statistică (§ 85), iar în al doilea, determinarea acestui interval se face cu ajutorul a doi parametri numai: valoarea medie şi abaterea tip (42).

Un alt rezultat important este faptul că, numărul elementelor dintr-un eşantion fiind mereu "suficient de mare", *valoarea medie a distribuţiei unei valori statistice este însăşi valoarea teoretică a acestei valori.*

Cu alte cuvinte, extrăgând un număr foarte mare de eşantioane dintr-o colectivitate care are înălţimea medie egală cu 1,68 m, valoarea mediei aritmetice a mediilor aritmetice ale eşantioanelor va fi egală cu 1,68.

Aşadar, media aritmetică a distribuţiei unei valori statistice fiind formal cunoscută, determinând şi abaterea-tip corespunzătoare acesteia, problema va fi complet rezolvată.

---

[1] *Iată de ce spuneam la începutul capitolului precedent că legea normală ne interesează în mod special în statistica matematică.*

Scanned with OKEN Scanner

Să facem o aplicație a considerațiilor precedente și, în acest scop, să ne referim la media aritmetică a unui eșantion considerată ca variabilă întâmplătoare. Știm că media aritmetică a acestei variabile întâmplătoare este media aritmetică a colectivității din care s-a extras eșantionul. Abaterea tip a ei este:

$$\sigma_{M'} = \frac{\sigma}{\sqrt{N}} \quad [1] \qquad (61)$$

unde σ este abaterea-tip a colectivității primitive, iar N este numărul elementelor din eșantion. Dacă acum M și M' sunt respectiv mediile aritmetice: a colectivității și a eșantionului, în baza proprietății legii normale, vom avea în general:

$$M - 2\frac{\sigma}{\sqrt{N}} \leq M' \leq M + 2\frac{\sigma}{\sqrt{N}}. \qquad (62)$$

În practică, de fapt, problema se pune în mod invers: M' este cunoscut și se caută să se determine un interval de variație a lui M. În afară de aceasta, σ este necunoscut. Se adoptă atunci, pentru o evaluare a lui, *abaterea-tip a eșantionului,* ceea ce înseamnă că:

$$\sigma_{M'} = \frac{\sigma'}{\sqrt{N}}. \qquad (61 \text{ bis})$$

Relația (62) devine, după transformare:

$$M' - 2\frac{\sigma'}{\sqrt{N}} \leq M \leq M' + 2\frac{\sigma'}{\sqrt{N}}. \qquad (62 \text{ bis})$$

Prin aplicarea ei, putem determina numai cu ajutorul valorilor eșantionului un interval în afara căruia media aritmetică a colectivității nu se poate găsi decât în mod cu totul excepțional.

Astfel, considerând eșantionul format din înălțimile celor 182 de tije de orz din tabelul 67, vom avea:

$$\sigma_{M'} = \frac{13{,}1422}{\sqrt{182}} = 0{,}974$$

și deci înălțimea medie teoretică a tijelor de orz de specia considerată va fi cuprinsă între 88,88 și 92,77.

Valoarea mediei teoretice n-a putut fi determinată exact, totuși am putea fixa un interval în care este cuprinsă, fapt care este departe de a fi comparat cu ignoranța completă a acesteia. Este una din formele practice sub care apar legile hazardului.

---

[1] *Formula aceasta se stabilește prin procedee expuse mai târziu* (Apendice).

103. *Forme concrete ale legii eşantioanelor numeroase.* Am văzut că, în ceea ce priveşte distribuţia normală, raportul între abaterea tip şi abaterea interquartilă este constant (§ 85). Am mai văzut, pe de altă parte, că există probabilităţi egale ca o abatere să fie mai mică sau mai mare decât abaterea interquartilă. Abaterea interquartilă a distribuţiei teoretice a mediei aritmetice a unui eşantion fiind:

$$Q_{M'} = 0{,}6745\ \sigma_{M'} = 0{,}6745\frac{\sigma}{\sqrt{N}} \qquad (63)$$

vom putea spune, ţinând seama de proprietatea abaterii interquartile, că există o probabilitate egală (şi egală deci cu ½) ca valoarea mediei unui eşantion să fie cuprinsă în intervalul $M - Q_{M'}$, $M + Q_{M'}$ sau să fie în afara lui. Inversând acum chestiunea, vom spune că probabilitatea ca să avem:

$$M' - Q_{M'} \leq M \leq M' + Q_{M'}$$

este egală cu aceea ca M să fie în afara acestui interval. Din această cauză, se obişnuieşte să se scrie valoarea mediei aritmetice dedusă dintr-un eşantion sub forma $M' \pm Q_{M'}$ unde:

$$Q_{M'} = 0{,}6745\frac{\sigma'}{\sqrt{N}} \qquad (63\text{ bis})$$

σ' fiind abaterea-tip a eşantionului.

În cazul distribuţiei înălţimii tijelor de orz, menţionate mai sus, vom avea:

$$M = 90{,}824 \pm 0{,}657.$$

Cunoaşterea abaterilor-tip sau a abaterilor interquartile ale diferitelor valori caracteristice ale unui eşantion, în cazul când N este suficient de mare, este aşadar de cea mai mare importanţă. Tabelul 82 cuprinde o listă a acestor abateri pentru valorile statistice cele mai uzitate.

Referindu-ne la tabelul 67, pentru a face o singură aplicaţie, vom avea:

$$\sigma = 13{,}1422 \pm 0{,}4645$$

precum şi:

$$\beta_2 = 2{,}5334 \pm 0{,}2450.$$

Această din urmă egalitate ne arată, ţinând seama de mărimea intervalului de variaţie statistică a lui β_2, că ipoteza făcută asupra normalităţii distribuţiei normale a înălţimii tijelor de orz din seria considerată nu constituie o absurditate: $3 \times 0{,}2450 = 0{,}7350$, β_2, putând varia, prin urmare, între 1,7984 şi 3,2784, ia în mod normal valoarea 3.

Scanned with OKEN Scanner

Se întâmplă contrariul în cazul distribuției din tabelul 66, unde:

$$\beta_2 = 3{,}442 \pm 0{,}0694$$

și deci intervalul de variație statistică a lui β_2 fiind 3,234 - 3,650, aceasta nu poate lua în mod acceptabil valoarea 3. Prin urmare, suntem îndreptățiți să spunem că mai degrabă distribuția considerată nu este un eșantion al unei colectivități normale, decât că este. Iată o concluzie care se apropie de cele stabilite mai sus referitor la aceeași distribuție (§ 88, 94).

TABELUL 82

Abaterile interquartile corespunzătoare celor mai importante constante statistice

Constanta statistică	Abaterea interquartilă
Colectivitatea primitivă oarecare	
Frecvențe relative	$0{,}6745\sqrt{\frac{pq}{N}}$
Media aritmetică	$0{,}6745\frac{\sigma}{\sqrt{N}}$
Fluctuația	$0{,}6745\sqrt{\frac{\mu_4 - \mu_2^2}{N}}$
Colectivitatea primitivă normală	
Mediana	$0{,}8454\frac{\sigma}{\sqrt{N}}$
Abaterea-tip	$0{,}4769\frac{\sigma}{\sqrt{N}}$
Quartil	$0{,}9191\frac{\sigma}{\sqrt{N}}$
Abaterea interquartilă	$0{,}5306\frac{\sigma}{\sqrt{N}}$
Coeficientul de variabilitate	$0{,}4769\frac{C.V.}{\sqrt{N}}\left[1+2\left(\frac{C.V.}{100}\right)^2\right]^{\frac{1}{2}}$
Fluctuația μ_2	$0{,}6745\sigma^2\sqrt{\frac{2}{N}}$
Momentul μ_3	$0{,}6745\sigma^3\sqrt{\frac{6}{N}}$

Scanned with OKEN Scanner

Momentul μ_4	$0{,}6745\sigma^4\sqrt{\frac{96}{N}}$
Coeficienţi pearsonieni $\sqrt{\beta_1}$	$0{,}6745\sqrt{\frac{6}{N}}$
Coeficienţi pearsonieni $\sqrt{\beta_2}$	$0{,}6745\sqrt{\frac{24}{N}}$

104. *Analiza mecanismului care conduce la diferitele forme de distribuţie*. În sfârşit, consideraţiile precedente asupra manifestărilor hazardului pot fi completate printr-o analiză a *formei* distribuţiilor heterograde. Am văzut mai sus (tabelul 79) că distribuţia numărului de cazuri în care se produce un eveniment de probabilitate p poate fi, după valorile acestei probabilităţi, normală, uşor şi pronunţat asimetrică sau în formă de J. Pornind de la acest rezultat, vom putea ajunge la o interpretare a formei distribuţiilor în legătură cu hazardul.

Să presupunem că toţi indivizii aparţinând aceleiaşi rase ar avea, în mod teoretic, o talie egală cu 1,65 m. Această egalitate de talii nu poate fi însă niciodată realizată, căci un ansamblu foarte numeros de cauze determină, prin acţiunea lor pentru fiecare caz în parte, o abatere de la talia ideală. Să mai presupunem că acţiunea unei astfel de cauze poate avea ca efect fie o micşorare a taliei cu 0,5 cm, fie o mărire a acesteia cu aceeaşi cantitate şi că, într-o masă foarte numeroasă de indivizi, abaterile în plus sau în minus de la talia ideală, datorate aceleiaşi cauze perturbatoare, sunt la fel de numeroase. Aceasta înseamnă că probabilitatea ca efectul unei astfel de cauze să fie o mărire sau o micşorare a taliei ideale este egală cu $\frac{1}{2}$. Admiţând valabilitatea acestei ipoteze, determinarea taliei efective a indivizilor din colectivitatea considerată se poate reprezenta prin schema următoare:

Dintr-o urnă, conţinând un număr foarte mare de bile albe şi negre, în egală proporţie, se scoate pentru fiecare individ şi la întâmplare un număr de bile egal cu numărul cauzelor care pot influenţa determinarea taliei. Să presupunem, pentru simplificarea expunerii, că acest număr este egal cu 12. Atunci cazurilor posibile ale frecvenţei bilei albe în 12 extracţii: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 le vor corespunde respectiv taliile: 1,59; 1,60; 1,61; 1,62; 1,63; 1,64; 1,65; 1,66; 1,67; 1,68; 1,69; 1,70; 1,71.

În acest mod se stabileşte o corespondenţă între distribuţia taliilor şi distribuţia frecvenţelor unei bile albe într-un eşantion de 12 extracţii dintr-o colectivitate în care bila albă are frecvenţa $p = \frac{1}{2}$. După cele ce ştim despre distribuţia binomială, vom deduce *a priori* că distribuţia taliilor într-o colectivitate omogenă este normală, cu atât mai mult cu cât, evident, numărul de cauze care

influențează determinarea taliilor și care corespunde numărului de extracții este în realitate extrem de numeros.

De fapt, problema considerată se pune exact în mod invers: se constată că taliile indivizilor de aceeași rasă se distribuie normal și atunci considerațiile precedente alcătuiesc o explicație plauzibilă a constatării făcute.

Biometricianul englez Sir FRANCIS GALTON a imaginat un aparat care să evidențieze în mod intuitiv faptul că *distribuția normală este rezultatul acțiunii unui număr foarte numeros de cauze acționând cu aceeași probabilitate când într-un sens, când într-altul.*

Aparatul constă dintr-o cutie; una din fețe este formată din sticlă, în scopul de a permite urmărirea vizuală a operației la care servește aparatul (graficul 42 A).

Prin pâlnia *P* se introduce în aparat o serie de alice. Aceste alice întâlnesc în cădere obstacolele *a*, care le fac să devieze în aceeași proporție când spre dreapta, când spre stânga și care reprezintă acțiunea peturbatoare. Bilele cad la sfârșit în niște despărțituri din fundul vasului, despărțituri care corespund intervalelor scării distribuției heterograde. Dacă numărul de alice introduse în vas este foarte mare, atunci se poate constata vizual că linia nivelului depozitelor de alice din diferitele despărțituri ale aparatului are forma curbei normale.

Prof. K. PEARSON, pornind de la ideea lui Galton, a perfecționat acest aparat, astfel ca operația descrisă să poată conduce și la formele asimetrice ale distribuției binomiale. În acest scop, obstacolele *a* sunt astfel plasate unele sub altele încât fiecare obstacol să împartă curentul alicelor ce se lovesc de el în raportul $q : p$. Referindu-ne la graficul 42 B, primul curent de N alice este împărțit de 1 în alte două conținând fiecare Nq și Np alice. Acum curentul de Nq alice este împărțit de obstacolul 2 în două: Nq^2 și Npq, iar cel de Np, de obstacolul 3 în: Npq și Np^2. Adică cele trei curente formate din alicele ce au întâlnit și al doilea rând de obstacole vor conține respectiv:

$$Nq^2 \qquad 2Npq \qquad Np^2$$

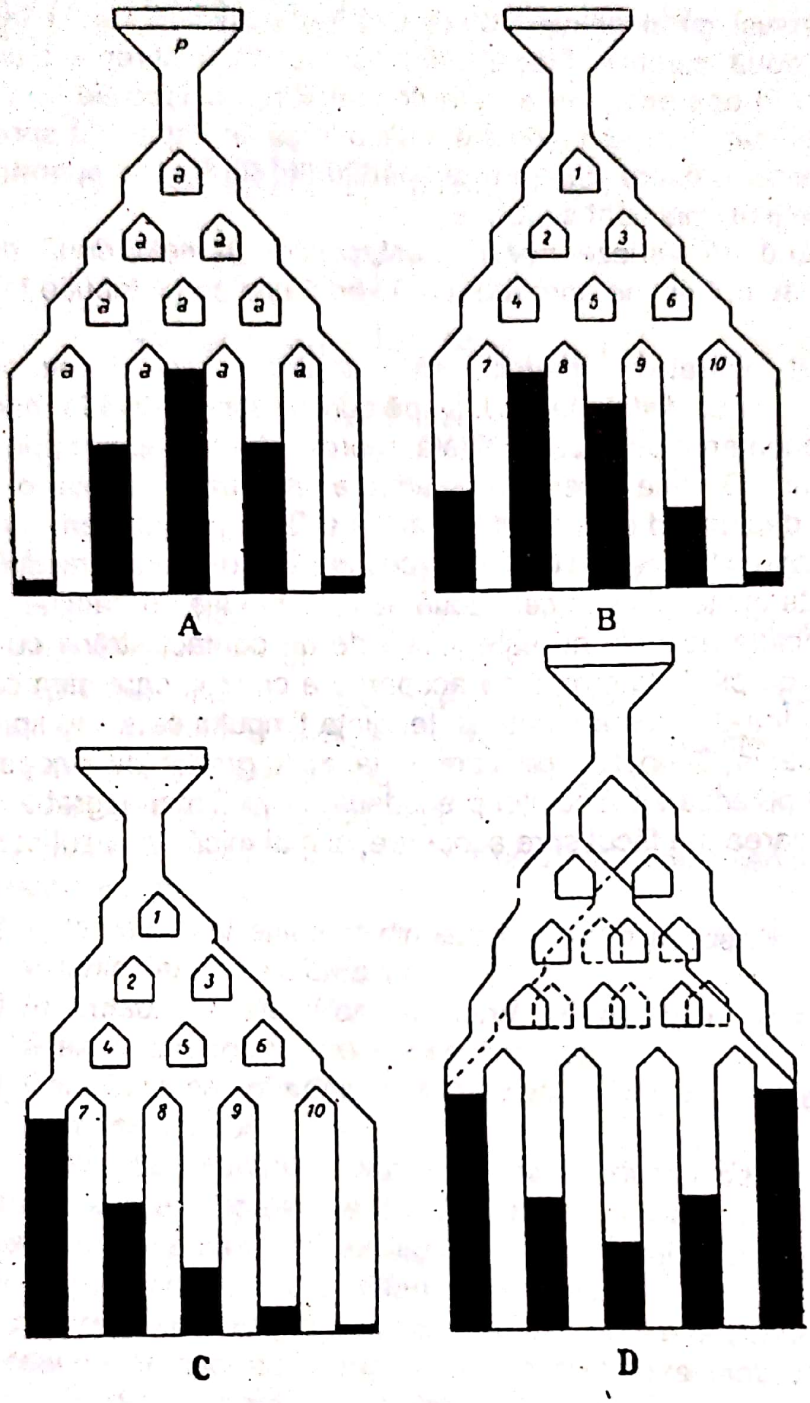
alice. Raționamentul se poate repeta; se ajunge astfel la repartiția din despărțiturile din fundul aparatului:

$$Nq^4,\ 4Nq^3p,\ 6Nq^2p^2,\ 4Nqp^3,\ Np^4.$$

Dacă obstacolele sunt aranjate ca în figura C, atunci obținem distribuția în J.

În legătură cu această construcție mecanică a distribuțiilor, putem obține și o interpretare a valorii medii. Dacă scoatem obstacolele din aparat, atunci toate alicele vor cădea în despărțitura din dreptul pâlniei. Această despărțitură reprezintă valoarea medie a distribuției considerate. Toate elementele ar fi egale cu această valoare dacă n-ar exista obstacolele, adică diferitele cauze ce produc abaterile de la valoarea medie.

Scanned with OKEN Scanner



Graficul 42. Aparatul Galton-Pearson pentru demonstrarea mecanismului formării distribuţiilor.

105. *Cazul distribuţiei în U.* Cu ajutorul aparatului Galton-Pearson se poate obţine orice tip de distribuţie dorim, mai ales dacă obstacolele nu mai sunt aranjate echidistant în interior, cu excepţia distribuţiilor în U.

Iată cum s-ar putea construi un aparat care să conducă la o astfel de distribuţie.

Scanned with OKEN Scanner

În însuşi gâtul pâlniei există un obstacol care de la început împarte alicele în două curente. Fiecare din aceste două curente ajunge în câte o despărţitură a aparatului. În aceste despărţituri, obstacolele sunt plasate astfel ca, în primul caz, marea majoritate a alicelor să fie deplasată spre stânga, iar în al doilea, spre dreapta. Singure despărţiturile din fundul aparatului în care se strâng alicele nu mai sunt separate.

Făcând să varieze poziţia obstacolelor în cele două despărţituri ale aparatului, se pot obţine aproape toate varietăţile de distribuţie în U (graficul 42 D).

Acest aparat ne permite să examinăm mai de aproape structura fenomenelor ce se distribuie în U. După cum vedem, există în manifestările unui astfel de fenomen o polarizare iniţială, reprezentată în aparat prin obstacolul din gâtul pâlniei. O dată această polarizare efectuată, există o tendinţă bine marcată şi depinzând de sensul în care s-a făcut polarizarea, ca fenomenul să capete un anumit aspect. Tendinţei acesteia îi corespunde modul diferit în care sunt aşezate obstacolele în cele două despărţituri ale aparatului.

Explicaţia aceasta nu este lipsită de un contact strâns cu realitatea. Să ne referim, de pildă, la gradul de acoperire a cerului, care ştim că se distribuie în U (§ 39). Într-un moment anumit, tendinţa timpului este sau spre o acoperire, sau spre senin. Valoarea pe care o ia apoi gradul de acoperire a cerului depinde de polarizarea tendinţei precedente şi va fi mai degrabă mare, în cazul când polarizarea s-a făcut spre acoperire, şi mai mică, în cazul contrar.

Scanned with OKEN Scanner

Capitolul VIII

Analiza seriilor homograde
Schemele lui Bernoulli, Poisson, Lexis

106. *Analiza eșantioanelor din punct de vedere al condițiilor în care sunt obținute*. Capitolul precedent ne-a arătat în ce măsură manifestările hazardului pot fi exprimate prin legi și cum ne putem servi de aceste legi în ceea ce privește aplicațiile practice ale statisticii matematice. Pentru motivul ca expunerea să permită o vedere de ansamblu a acestei chestiuni așa de delicate, am lăsat înadins deoparte discuția asupra felului cum este format eșantionul ce constituie seria statistică.

Cel mai simplu caz de studiat se prezintă când eșantionul îndeplinește următoarele condiții:

a) *Valoarea unui element din eșantion este pe deplin independentă de valorile celorlalte elemente.*
b) *În cursul observațiilor alcătuind un eșantion al unei colectivități, aceasta din urmă rămâne neschimbată.*
c) *Colectivitatea rămâne neschimbată chiar dacă trecem de la un eșantion la altul.*

Un astfel de eșantion *ideal* se prezintă destul de rar în practică, de aceea mai toate exemplele ce se dau pentru ilustrarea lui sunt artificiale: aruncarea banului, a zarurilor, extragerea unei bile albe dintr-o urnă conținând un număr foarte mare de bile de diferite culori etc.

Înlăturarea uneia sau a două condiții din cele menționate ne conduce la tipuri de eșantioane care au un mai strâns contact cu realitatea; studiul acestora este însă mai dificil decât al eșantionului *ideal*. Bineînțeles că înlăturarea tuturor condițiilor precedente ne conduce la un eșantion care nu mai are nimic tipic în el și care, prin urmare, nu mai poate servi scopului analitic al statisticii matematice. Vom da câteva exemple de eșantioane care nu mai îndeplinesc decât o parte din condițiile *a, b, c*.

Frecvența bolnavilor de o boală contagioasă nu îndeplinește condiția *a*, deoarece fiecare bolnav, prin focarul de contagiune pe care-l constituie,

Scanned with OKEN Scanner

mărește șansa pe care o au ceilalți indivizi, încă sănătoși, de a se îmbolnăvi și ei. Același lucru se poate spune și despre numărul morților și răniților în accidentele de tren. Existența unui rănit sau mort într-un astfel de accident înseamnă că accidentul s-a produs și că, prin urmare, există toate șansele ca să mai fie și alte persoane în aceeași categorie.

În ceea ce privește condiția *b*, putem afirma că realitatea nu ne oferă decât cazuri când aceasta nu este îndeplinită. Este într-adevăr efectiv imposibil ca să determinăm o colectivitate perfect omogenă, care să nu fie compusă, prin urmare, din grupuri diferite. Frecvența născuților de sex masculin față de numărul total al nașterilor, de pildă, este alcătuită din două eșantioane: primul relativ la născuții legitimi, cel de-al doilea, la născuții nelegitimi. Separarea aceasta nu este însă suficientă, deoarece în interiorul acestor clase se mai pot face diferențieri în raport cu vârsta tatălui și a mamei, religia lor, profesiunea acestora, numărul de ani scurși de la primul născut al mamei, numărul de copii născuți înaintea celui considerat etc. Cu toate acestea, trebuie să recunoaștem că nu toate atributele care pot diferenția elementele unei colectivități au aceeași influență asupra fenomenului urmărit. Căutând să diferențiem colectivitatea în raport cu acele atribute care au o influență vădită asupra fenomenului studiat, vom obține colectivități *practic* omogene din punctul de vedere al problemei puse. În cazul coeficientului de masculinitate, este indicat să se țină seamă de variația vârstei tatălui și a mamei. Bineînțeles însă că s-ar putea ca o astfel de colectivitate să nu mai fie omogenă pentru o altă problemă.

Eliminând condiția *c*, am putea avea impresia, din cauza structurii ei, că orice punct de sprijin al metodei statistice este pierdut. Într-adevăr, ce concluzii s-ar putea trage din studiul unei serii demografice care prima dată s-ar referi la populația unui oraș din România, a doua la un oraș din Franța, a treia oară la unul din Chile etc.? Evident că nici o concluzie, afară numai dacă fenomenul considerat nu este supus unei *legi numerice universale.* Problema însă nu se pune în acest mod; se înlătură condiția *c* fără ca să permitem colectivităților din care extragem eșantioanele să se deosebească prea mult una de alta. Mortalitatea în diferitele județe ale României formează o serie statistică satisfăcând aceste noi condiții.

107. *Schema Bernoulli. Coeficientul de dispersie lexian.* Vom trece acum să examinăm formele concrete sub care ni se înfățișează diferitele tipuri de eșantioane considerate, precum și mijloacele practice de care dispunem ca să le recunoaștem[1]. Ne vom mărgini la studiul frecvențelor producerii unui eveniment oarecare, adică la cel al seriilor homograde.

Vom începe cu cazul eșantionului ideal. Acesta a fost studiat în primul rând de matematicianul JACQUES BERNOULLI, care l-a abstractizat printr-o schemă ce-i poartă numele. Schema lui Bernoulli constă din extracții repetate făcute dintr-o urnă care conține un număr foarte mare de bile albe și negre în

[1] *Nu vom examina cazul eșantioanelor ce nu îndeplinesc condiția a, caz care este cel mai complicat. Statistica matematică nu posedă încă în această direcție rezultatele precise, care se cunosc de mult în cazul independenței de probabilități.*

Scanned with OKEN Scanner

raportul $p : q$. Cazul acesta a fost examinat complet în capitolul precedent. Să rezumăm:

Se fac n extracţii, se repetă această operaţie de s ori, obţinându-se astfel frecvenţele statistice:

$$f_1 = \frac{n_1}{n}, \quad f_2 = \frac{n_2}{n}, \ldots, \quad f_s = \frac{n_s}{n}.$$

Media aritmetică şi abaterea-tip ale acestei serii au valorile teoretice:

$$M = p \qquad \sigma_B = \sqrt{\frac{pq}{n}}.$$

Fiecare din frecvenţele f este o evaluare statistică a lui p cu dispersia menţionată. O evaluare şi mai bună a mediei teoretice este:

$$p_0 = \frac{f_1 + f_2 + \ldots + f_s}{s} \tag{64}$$

a cărei dispersie va fi evident măsurată de abaterea-tip:

$$\sigma_{p_0} = \sqrt{\frac{pq}{sn}} = \frac{\sigma_B}{\sqrt{s}}. \tag{65}$$

O evaluare a abaterii-tip a frecvenţei f va fi dată de relaţia:

$$\sigma'_B = \sqrt{\frac{p_0 q_0}{n}}. \tag{66}$$

Aceasta se poate calcula *numai cu ajutorul valorii mediei* p_0, pentru care motiv o vom şi numi *valoare calculată*, spre deosebire de *valoarea măsurată* a aceleiaşi abateri, care va fi dată de relaţia:

$$\sigma_f^2 = \frac{\Sigma(f_i - p_0)^2}{s}\ ^1. \tag{67}$$

Dacă frecvenţele f provin într-adevăr dintr-o schemă Bernoulli, atunci coeficientul de dispersie introdus de LEXIS[2]*:*

---

[1] *În cazul când s este prea mic, este recomandabil să se ia:*

$$\sigma_f^2 = \frac{\Sigma(f_i - p_0)^2}{s - 1}.$$

[2] *W. Lexis,* ***Zur Theorie der Massenerscheinungen in der Menschlichen Gesellschaft****, Fribourg-en-Brisgau, 1877.*

Scanned with OKEN Scanner

$$L = \frac{\sigma_f}{\sigma_B} \tag{68}$$

va fi teoretic egal cu unitatea, iar în practică, vecin acesteia.

108. *Cazul eşantioanelor formate dintr-un număr diferit de observaţii.* În practică s-ar putea întâmpla ca şirul celor s eşantioane să nu conţină acelaşi număr de observaţii, adică să avem:

$$f_1 = \frac{n_1}{m_1}, \quad f_2 = \frac{n_2}{m_2}, \ldots, f_s = \frac{n_s}{m_s}.$$

În acest caz, p_0 va fi o *medie ponderată* a valorilor f:

$$p_0 = \frac{\Sigma mf}{\Sigma m} = \frac{\Sigma n}{\Sigma m} \tag{64 bis}$$

a cărei dispersie va fi:

$$\sigma_{p_0} = \sqrt{\frac{pq}{\Sigma m}} \tag{65 bis}$$

valoare care se reduce la (65) când toate valorile m sunt egale între ele.

La fel, vom avea:

$$\sigma'_B = \sqrt{\frac{sp_0q_0}{\Sigma m}} \tag{66 bis}$$

şi:

$$\sigma_f^2 = \frac{\Sigma m(f_i - p_0)^2}{\Sigma m}\ ^1. \tag{67 bis}$$

Vom avea însă foarte rar ocazia să aplicăm aceste formule, deoarece valorile m, în mai toate cazurile de serii statistice, nu diferă prea mult una de alta şi, din această cauză, diferenţa între mediile simple şi cele ponderate nu va fi aşa de mare încât să necesite întrebuinţarea acestora din urmă.

109. *Schema Bernoulli şi naşterile masculine.* Să aplicăm cele de mai sus studiului coeficientului de masculinitate al naşterilor din Bucureşti pentru perioada 1886-1915 (tabelul 83). Datele care au servit ca bază calculelor din acest tabel au

---

[1] *Când s e mic,*

$$\sigma_f^2 = \frac{s}{s-1} \cdot \frac{\Sigma m(f_i - p_0)^2}{\Sigma m}.$$

Scanned with OKEN Scanner

fost extrase din G. *Milea*, ***Contribuţii la demografia municipiului Bucureşti***, p. 30-31[1].

TABELUL 83

Calculul coeficientului lui Lexis pentru coeficientul de masculinitate al naşterilor din Bucureşti (1886-1915)

Anul	Naşteri		Coeficient de masculinitate	Abateri faţă de 0,5126	Abateri la pătrat
	băieţi	total			
1886	3 296	6 592	0,5000	126	15 876
1887	3 305	6 453	122	4	16
1888	3 497	6 753	178	52	2 704
1889	3 432	6 676	141	15	225
1890	3 360	6 649	053	73	5 329
1891	3 724	7 214	162	36	1 296
1892	3 479	6 734	166	40	1 600
1893	3 838	7 458	146	20	400
1894	3 909	7 680	090	36	1 296
1895	4 260	8 170	214	88	7 744
1896	4 140	8 112	104	22	484
1897	4 584	8 949	122	4	16
1898	4 262	8 462	037	89	7 921
1899	4 696	9 245	075	51	2 601
1900	4 481	8 766	112	14	196
1901	4 409	8 547	159	33	1 089
1902	4 207	8 091	200	74	5 476
1903	4 093	7 954	146	20	400
1904	4 239	8 167	190	64	4 096
1905	4 136	7 915	226	100	10 000
1906	4 383	8 448	188	62	3 844
1907	4 442	8 647	137	11	121
1908	4 757	9 156	196	70	4 900
1909	4 704	9 089	175	49	2 401
1910	4 627	9 151	056	70	4 900
1911	4 865	9 619	058	68	4 624
1912	5 066	9 976	078	48	2 304
1913	5 146	9 975	159	33	1 089
1914	5 334	10 485	087	39	1 521
1915	5 317	10 601	016	110	21 100
	127 988	249 734	15,3793	-	106 569

[1] *Lucrare prezentată Şcolii de Statistică pentru obţinerea diplomei de statistician, 1932.*

Scanned with OKEN Scanner

Să calculăm valorile mediei aritmetice și ale abaterilor tip calculate și măsurate, făcând presupunerea că fiecare eșantion cuprinde același număr de elemente, pe care-l vom lua egal cu media aritmetică a numărului anual de nașteri.

$$\frac{\Sigma m}{s} = \frac{249734}{30} = 8324.$$

Vom avea în acest caz:

$$p_0 = \frac{\Sigma f}{s} = \frac{15{,}3793}{30} = 0{,}5126$$

precum și:

$$\sigma_f = \sqrt{\frac{0{,}00106569}{30}} = 0{,}00596 \qquad \sigma'_B = \sqrt{\frac{0{,}5126 \times 0{,}4874}{8324}} = 0{,}00548$$

și deci:

$$L = \frac{0{,}00596}{0{,}00548} = 1{,}088.$$

Această valoare a lui L diferă așa de puțin de unitate, încât avem tot dreptul să afirmăm că seria studiată se calculează pe o schemă Bernoulli.

Dacă vrem să ținem seamă de faptul că nu s-a născut același număr de copii în toți anii, vom întrebuința formulele din § 108. Astfel, vom avea:

$$p_0 = \frac{\Sigma n}{\Sigma m} = \frac{127988}{249734} = 0{,}5125 \qquad \sigma_f = \sqrt{\frac{\Sigma m (f - p_0)^2}{\Sigma m}} = 0{,}00597$$

$$\sigma'_B = \sqrt{\frac{30 \times 0{,}5125 \times 0{,}4875}{249734}} = 0{,}00548 \ [1].$$

[1] *Egalitatea aceasta se mai poate scrie:*

$$\sigma'_B = \sqrt{\frac{0{,}5125 \times 0{,}4875}{8324}}$$

ceea ce înseamnă că σ'_B se obține ca mai sus, adică atribuind fiecărui eșantion un număr de elemente egal cu media aritmetică a acestora: $\frac{\Sigma m}{s}$.

Analiza mai aprofundată a chestiunii ne arată că logic ar fi să atribuim eșantioanelor un număr de elemente egal cu media armonică *a valorilor m. (Ch. Jordan,* ***Statistique mathématique****, Paris, Gauthier-Villars, 1927, p. 196, U. G. Yule,* ***op. cit.****, p.264). Ținând seamă că acest procedeu nu se poate aplica decât în cazul când valorile m nu diferă prea mult între ele, când, prin urmare, media este sensibil egală cu media aritmetică și că media armonică necesită un calcul prea laborios, aș înclina pentru întrebuințarea practică a mediei aritmetice pentru acest scop, mai ales în cazul când eșantioanele cuprind un număr mai mare de elemente.*

Scanned with OKEN Scanner

Se verifică astfel faptul anunțat dinainte că formulele care țin seama de ponderare conduc în practică la rezultate sensibil egale cu cele în care nu intervine ponderarea.

110. *Schema hiponormală.* POISSON a construit o schemă diferită de precedenta, prin înlăturarea condiției *b*, (§ 106). În acest caz, fiecare eșantion compus din *m* elemente va fi obținut din *m* extracții făcute în *m* colectivități diferite, C_1, C_2, ..., C_m, care însă rămân neschimbate pentru toate eșantioanele.

Dacă p_1, p_2, ..., p_m sunt frecvențele cazului urmărit în cele *m* colectivități, atunci, ignorând existența acestora, vom atribui unei colectivități unice, din care s-ar extrage toate elementele eșantioanelor, frecvența:

$$p = \frac{p_1 + p_2 + \ldots + p_m}{m}.$$

Vom avea o evaluare a acestei frecvențe prin frecvențele statistice:

$$p_0 = \frac{f_1 + f_2 + \ldots + f_s}{s}$$

a cărei dispersie va fi dată de abaterea-tip:

$$\sigma_{p_0} = \frac{1}{m}\sqrt{\frac{\Sigma p_i q_i}{s}}. \qquad (69)$$

Trecând mai departe și calculând abaterile-tip calculate și măsurate ale frecvențelor *f*, vom avea:

$$\sigma'_B = \sqrt{\frac{p_0 q_0}{m}}$$

$$\sigma_f^2 = \frac{\Sigma (f - p_0)^2}{s}\ ^1.$$

Ultima abatere-tip are ca valoare teoretică:

$$\sigma_f^2 = \frac{\Sigma p_i q_i}{m^2} = \frac{pq}{m} - \frac{\Sigma (p_i - p)^2}{m^2} = \sigma_B^2 - \frac{\Sigma (p_i - p)^2}{m^2}$$

deci:

$$\sigma_f < \sigma_B$$

[1] *Când s este prea mic, atunci numitorul se înlocuiește cu* s - 1.

și, prin urmare:

$$L < 1 . \tag{70}$$

Acesta este simptomul prin care putem recunoaște în mod practic dacă o serie corespunde sau nu schemei Poisson. Termenii unei astfel de serii prezintă o dispersie σ_f mai mică decât ar fi rezultat dintr-o schemă Bernoulli, fapt care explică termenul de *serie hiponormală* care se întrebuințează pentru aceste distribuții.

Este important de reținut faptul că dispersia unei serii ai cărei termeni sunt rezultatul unor extrageri din colectivități diferite este mai mică decât dacă eșantioanele ar fi reprezentat o colectivitate unică: frecvențele numerelor pare vor avea o dispersie mai mică în cazul când eșantionul este obținut dintr-o singură aruncătură de 12 *zaruri diferite* și mai mare în cazul unui eșantion format din 12 aruncături ale *aceluiași zar*, mai ales dacă zarurile nu sunt prea îngrijit construite. *Abaterile rezultând din cauze diferite tind să se compenseze.*

111. *Schema hipernormală.* Schema lui LEXIS, care i-a prilejuit de altfel introducerea coeficientului L, corespunde eșantioanelor pentru care condiția c nu mai este îndeplinită. Primul eșantion este extras dintr-o colectivitate în care frecvența este p_1, al doilea dintr-o colectivitate în care frecvența este p_2 etc. Ignorând acest fapt, vom asimila aceste extracții cu eșantioane ale aceleiași colectivități pentru care frecvența cazului urmărit va fi:

$$p = \frac{p_1 + p_2 + \ldots + p_s}{s} .$$

O evaluare statistică a lui p va fi:

$$p_0 = \frac{f_1 + f_2 + \ldots + f_s}{s}$$

cu abaterea-tip:

$$\sigma_{p_0} = \frac{1}{s}\sqrt{\frac{\sum p_i q_i}{m}} . \tag{71}$$

Vom calcula abaterile-tip, calculate și măsurate, cu ajutorul formulelor (66), (67), eventual (66 bis), (67 bis). Însă în acest caz:

$$\sigma_f^2 = \frac{pq}{m} + \left(1 - \frac{1}{m}\right)\frac{\sum (p_i - p)^2}{s}$$

și deci $\sigma_f > \sigma_B$, prin urmare:

$$L > 1 . \tag{72}$$

Scanned with OKEN Scanner

Schema lui Lexis dă naştere la *serii hipernormale*. Acestea sunt cele care se întâlnesc cel mai des. În practică, este într-adevăr aproape imposibil să menţinem colectivitatea neschimbată de la un eşantion la altul. În majoritatea cazurilor, fenomenele colective sunt supuse unei variaţii în timp, care se traduce prin variaţia structurii colectivităţii din care extragem eşantioanele.

Acum se explică în parte de ce seriile hipernormale sunt extrem de rare în statistică, deşi la prima vedere ar trebui să ne aşteptăm să le întâlnim cel mai des: condiţiile care conduc la hipernormalitate se întretaie cu cele de hiponormalitate; primele fiind de obicei covârşitoare, obţinem în general distribuţii hipernormale sau, în cel mai bun caz, normale.

Un exemplu de schemă Lexis ni-l oferă distribuţia din tabelul 84.

TABELUL 84

Mortalitatea de tuberculoză pulmonară în România, 1906

Vârsta	Morţi de tuberculoză	Totalul morţilor	Frecvenţa morţilor de tuberculoză	Abateri faţă de 0,126	Abateri la pătrat	(3) × (6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sub 1 an	51	7 142	0,007	0,119	0,014161	101,137862
1 - 9 ani	233	3 873	0,060	0,066	0,004356	16,870788
10 - 19 ani	579	1 522	0,380	0,254	0,064516	98,193352
20 - 39 ani	1 567	3 974	0,394	0,268	0,071824	285,428576
40 - 59 ani	899	4 703	0,191	0,065	0,004225	19,870175
Peste 60 ani	235	6 985	0,034	0,092	0,008464	59,121040
	3 564	28 199	1,066	-	0,167546	580,621793

Ţinând seamă de ponderi şi de corecţiile menţionate mai sus pentru cazul când *s* este mic, vom avea:

$$p_0 = \frac{3564}{28199} = 0{,}126$$

$$\sigma_f^2 = 1{,}2 \times \frac{580{,}621793}{28212} = 0{,}024696804 \qquad \sigma_f = 0{,}1572$$

$$\sigma_B = \sqrt{\frac{0{,}126 \times 0{,}874}{4702}} = 0{,}0048$$

$$L = 32{,}75\ ^{1}$$

ceea ce indică o hipernormalitate pronunțată: mortalitatea de tuberculoză pulmonară variază neîndoios cu vârsta.

112. *Coeficientul de instabilitate al lui Charlier.* Criteriile precedente se pot aplica întocmai și seriilor de frecvențe absolute. În acest caz:

$$n_0 = \frac{n_1 + n_2 + \ldots + n_s}{s} \qquad p_0 = n_0 m$$

$$\sigma'_B = \sqrt{m p_0 q_0}$$

$$\sigma_n^2 = \frac{\Sigma (n_i - n)_0^2}{s}$$

și coeficientul lui Lexis:

$$L = \frac{\sigma_n}{\sigma'_B}. \tag{73}$$

Valoarea lui L este influențată în acest caz de numărul elementelor din eșantion, tendința fiind ca să aibă valori mai mari cu cât m este mai mare. CHARLIER, căutând să evite această particularitate a coeficientului lexian, a introdus *coeficientul de instabilitate:*

$$C = \frac{100}{p_0} \sqrt{\sigma_f^2 - \sigma_B^2} \tag{74}$$

care este independent de numărul de observații din eșantion.

Vom avea o serie normală dacă $C = 0$, hiponormală când C este imaginar, iar C pozitiv va indica o distribuție hipernormală.

În cazul seriilor din tabelele 83 și 84, vom avea respectiv:

$$C = 0{,}46 \qquad C = 124{,}67.$$

---

[1] *Întrebuințând formulele simple (neponderate), obținem:*

$$p_0 = \frac{1{,}066}{6} = 0{,}178 \qquad \sigma_f^2 = \frac{0{,}167546}{6} - 0{,}052^2 = 0{,}025220 \ (necorectat)$$

$$\sigma_f^2 = 1{,}2 \times 0{,}025220 = 0{,}030264 \ (corectat) \qquad \sigma_f = 0{,}1740$$

$$\sigma_B = \sqrt{\frac{0{,}178 \times 0{,}822}{4702}} = 0{,}0056 \qquad L = 31{,}07$$

Scanned with OKEN Scanner

113. *Eliminarea tendinței evolutive a fenomenelor colective.* În ceea ce privește hipernormalitatea seriilor în timp, aceasta poate fi înlăturată prin nivelarea structurilor colectivităților observate. Să ne fixăm ideile asupra natalității din municipiul București, pentru perioada 1886-1910 (tabelul 85). O simplă inspecție a datelor ne arată că natalitatea considerată a descrescut simțitor în intervalul de timp considerat. Această descreștere s-a putut efectua după o lege foarte complicată, însă cea mai simplă ipoteză pe care o putem admite și care în același timp se calchează destul de bine pe majoritatea cazurilor este descreșterea (eventual creșterea) uniformă, adică aceeași de la an la an.

TABELUL 85

Eliminarea tendinței evolutive a coeficientului de natalitate din București (1886-1910)

(Populația medie = 260 360 locuitori)

Anul	Coeficient de natalitate	Abateri $f_i - p_0$	$i - \frac{s+1}{2}$	$\left(i - \frac{s+1}{2}\right) \times (f_i - p_0)$ (3) x (4)	$\left(i - \frac{s+1}{2}\right)\beta$	(2) + (6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1886	0,03624	+ 502	- 12	- 6 024	- 540	0,03084
1887	3531	+ 409	- 11	- 4 499	- 495	3036
1888	3682	+ 560	- 10	- 5 600	- 450	3232
1889	3618	+ 496	- 9	- 4 464	- 405	3213
1890	3428	+ 306	- 8	- 2 448	- 360	3068
1891	3547	+ 425	- 7	- 2 975	- 315	3232
1892	3174	+ 52	- 6	- 312	- 270	2904
1893	3363	+ 241	- 5	- 1 205	- 225	3138
1894	3310	+ 188	- 4	- 752	- 180	3130
1895	3371	+ 249	- 3	- 747	- 135	3236
1896	3218	+ 96	- 2	- 192	- 90	3128
1897	3410	+ 288	- 1	- 288	- 45	3365
1898	3110	- 12	0	-	-	3110
1899	3277	+ 155	+ 1	+ 155	+ 45	3322
1900	3052	- 70	+ 2	- 140	+ 90	3142
1901	2928	- 194	+ 3	- 582	+135	3063
1902	2730	- 392	+ 4	- 1 568	+180	2910
1903	2645	- 477	+ 5	- 2 385	+225	2870
1904	2676	- 446	+ 6	- 2 676	+270	2946
1905	2563	- 559	+ 7	- 3 913	+315	2878
1906	2694	- 428	+ 8	- 3 424	+360	3054
1907	2721	- 401	+ 9	- 3 609	+405	3126

Scanned with OKEN Scanner

(continuare tab. 85)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1908	2837	- 285	+ 10	- 2 850	+450	3287
1909	2781	- 341	+ 11	- 3 751	+495	3276
1910	2762	- 360	+ 12	- 4 320	+540	3302
	0,78052	-	-	- 58 724 + 155	-	
-	-	-	-	- 58 569	-	

$$p_0 = \frac{0{,}78052}{25} = 0{,}03122 \qquad \sigma'^2_B = \frac{0{,}3122 \times 0{,}6878}{260360} = 0{,}0000001162$$

$$\sigma_f^2 = 0{,}0000125354\,(col.2) \qquad \sigma_f^2 = 0{,}0000019806 \quad (col.7)$$

Valoarea teoretică a acestei creşteri este dată de formula:

$$\beta = \frac{12}{s(s^2-1)} \sum \left(i - \frac{s+1}{2} \right) (f_i - p_0). \tag{75}$$

Amănuntele de calcul ale aplicării acestei formule sunt arătate în tabelul 85. După calculul valorii β, datele liberate de mişcarea evolutivă a fenomenului se obţin din cele brute prin adăugarea termenului $\left(i - \frac{s+1}{2} \right) \beta$ [1] (coloana 7).

Rezultatele calculelor din tabelul menţionat ne conduc la :

$$\beta = \frac{12}{25(625-1)} \times (-0{,}58569) = -0{,}00045.$$

Coeficientul de instabilitate este:

$$C = 11{,}27$$

pentru seria din coloana (2), şi :

$$C = 4{,}39$$

pentru cea din coloana (7), ceea ce înseamnă că prin operaţia făcută s-au înlăturat în bună parte tendinţele care contribuiau la hipernormalitatea seriei considerate.

---

[1] *Media aritmetică a datelor va rămâne neschimbată, căci*

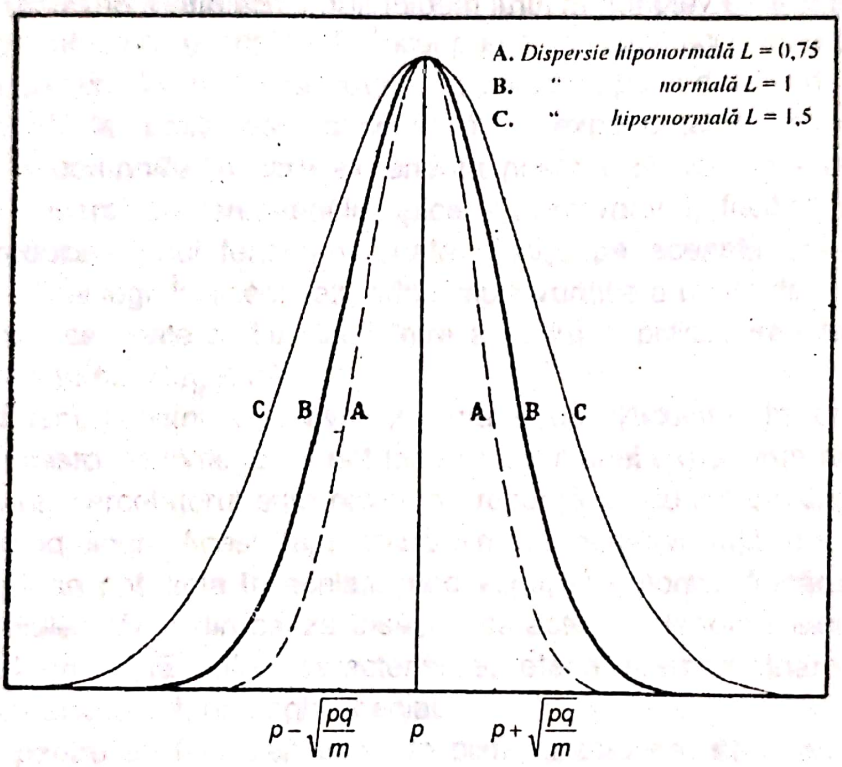
$$\sum \left(i - \frac{s+1}{2} \right) = 0.$$

Scanned with OKEN Scanner

Ultima valoare a coeficientului *C* este totuşi suficient de mare ca natalitatea din municipiul Bucureşti să nu poată fi privită ca un fenomen cu dispersie normală, ceea ce se poate explica prin faptul că acest oraş prezintă o populaţie cu o structură schimbată continuu de către elementul imigrator.

114. *Comparaţie între cele trei tipuri de distribuţii normale.* Cercetarea unei serii statistice din punct de vedere al *instabilităţii* poate să conducă la rezultate practice din cele mai interesante, mai ales atunci când aspectul seriei este interpretat simultan şi din punctul de vedere al diferenţelor semnificative.

Înainte de a trece la examinarea acestei noi probleme, lucru care-l vom face în capitolul următor, este necesar să accentuăm ceva mai mult natura diferenţei dintre structurile seriilor statistice normale, hipo- şi hipernormale.



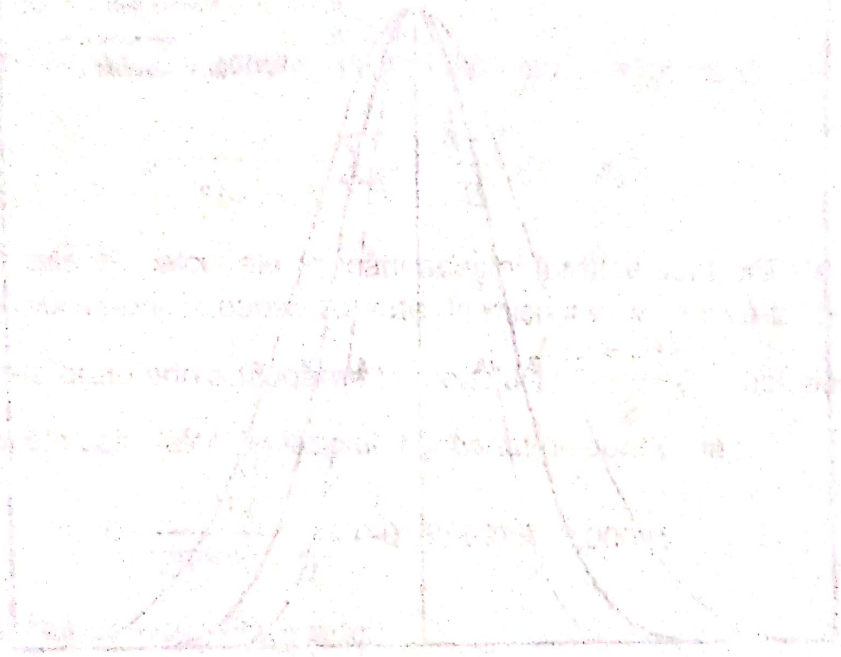
Graficul 43. Comparaţie între diferitele structuri ale curbei normale.

În cazul eşantioanelor numeroase, frecvenţele statistice se distribuie după legea normală, oricare ar fi schema corespunzătoare acestora. Legea normală a cărei ecuaţie este:

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{(x-M)^2}{2\sigma^2}}$$

depinde de două variabile: media aritmetică *M* şi abaterea-tip σ. Valorile acestor constante se pot calcula, după cum ştim, pornind de la seria statistică

respectivă, şi sunt în general independente una de alta[1]. În cazul special al frecvenţelor, se poate întâmpla ca valoarea abaterii-tip să fie egală cu $\sqrt{\frac{pq}{m}}$, când, prin urmare, poate fi determinată numai prin cunoaşterea mediei $M = p$. O astfel de serie se numeşte cu dispersie normală şi poate fi interpretată printr-o schemă Bernoulli. În celelalte cazuri avem sau o serie cu dispersie hiponormală, $\sigma^2 < \frac{pq}{m}$, sau hipernormală, $\sigma^2 > \frac{pq}{m}$. Termenul $\frac{pq}{m}$ n-are nici o legătură firească cu seriile ce nu sunt normale, el serveşte numai ca un punct util de comparaţie. Graficul 43 ne arată diferenţa ce există între trei serii a căror dispersie corespunde celor trei scheme studiate: forma curbelor este aceeaşi, singură concentrarea valorilor în jurul mediei aritmetice diferă de la caz la caz.



[1] *Independenţă fizică, nu independenţă în sensul probabilistic.*

Scanned with OKEN Scanner

Capitolul IX

Comparația statistică

115. *Caracterul comparației fenomenelor colective.* Scopul final al oricărei discipline științifice este formularea de legi, adică stabilirea unei legături între fenomene, între manifestările acestora sau pur și simplu între prezența sau absența unor caractere esențiale. Mijlocul prin care acest scop este atins diferă după natura domeniului în care se întreprinde studiul, putând varia de la inducția completă până la observația descriptivă a experiențelor provocate sau întâmplate. În domeniile în care experiența poate fi provocată - ceea ce se întâmplă în general cu fenomenele tipice - observatorul, făcând să varieze condițiile producerii unui fenomen, poate ajunge pe această cale numai la descoperirea unei legi. În acest caz, prin simpla variație a uneia din condițiile de experimentare, se poate stabili dacă între aceasta și producerea fenomenului studiat există sau nu o legătură.

Cazul fenomenelor colective prezintă două dificultăți. În primul rând, majoritatea acestor fenomene nu pot face obiectul unei experiențe de laborator și, prin urmare, cercetătorul este nevoit să recurgă la studiul cazurilor care se produc în mod liber. Acest fapt constituie un neajuns față de experiența provocată: nu se pot izola în același mod variațiile datorate fiecărei cauze în parte. În al doilea rând, din cauza însăși a caracterului fenomenelor colective, comparația între două valori caracteristice, etapa premergătoare formulării legilor, trebuie privită sub un unghi special.

Dacă pentru un fenomen tipic, de pildă, presiunea, se poate răspunde precis în care din două cazuri aceasta este mai mare, temperatura variind, nu vom putea ajunge tot așa de ușor să spunem care din două rase de oameni are o talie mai mare. În primul caz, diferența între două valori ale presiunii datorate variației temperaturii este întotdeauna *semnificativă*, în al doilea, diferența între două talii nu poate fi semnificativă pentru colectivitatea întreagă. Considerând la întâmplare talia unui individ de rasă albă, vom găsi negri a căror talie să fie mai mare decât prima, dar vom găsi de asemenea alții a căror talie să fie mai mică. Din această simplă comparație nu se poate ajunge la formularea nici unei legi. Este natural deci să renunțăm pentru studiul fenomenelor colective la comparația valorilor individuale ale colectivităților. Utilitatea valorilor tipice pentru cercetarea științifică a fost impusă aproape numai de această necesitate.

Scanned with OKEN Scanner

Nici comparaţia acestor valori nu este însă lipsită de părţi delicate, iar uneori spinoase. Revenim la exemplul de mai sus. Presupunând că am măsurat taliile a două grupuri - primul de rasă neagră, celălalt de rasă albă - şi că am găsit respectiv taliile medii egale cu 1,78 cm şi 1,68 cm, putem oare afirma că rasa neagră are o talie mai mare decât cea albă, fără ca afirmaţia noastră să poate fi atacată pe motivul unui viciu de raţionament? Diferenţa observată poate fi într-adevăr interpretată ca provenind din *individualitatea* celor două grupuri, căci înseşi grupurile au un caracter individual şi comparaţia lor este analoagă până la un punct cu comparaţia a doi indivizi izolaţi.

Iată cum se pune, în linii generale, problema comparaţiei statistice despre care se ocupă mai pe larg paragrafele următoare. De la început vom distinge două cazuri:

1. Comparaţia seriilor statistice între ele;
2. Comparaţia valorilor caracteristice.

116. *Comparaţia grafică directă*. Comparaţia seriilor statistice se poate face în două feluri: grafic şi prin calcul.

Ne vom ocupa în special de cazul distribuţiilor de frecvenţă[1]. Pentru a ne fixa ideile, să considerăm distribuţiile de salarii din tabelul 86.

În mod grafic, comparaţia acestor serii se poate face figurând pe acelaşi grafic poligoanele de frecvenţă respective (graficul 44 A).

Se constată astfel că salariile din statul A sunt mai mici decât cele din statul B, acuzând în acelaşi timp şi o dispersie mai mică. Comparaţia prin histograme este mai anevoioasă, dacă nu chiar imposibilă.

După cum vedem, comparaţia grafică nu ne poate conduce decât la rezultate de formă ceva cam simplistă. Este totuşi de recomandat ca această operaţie să se facă cu ajutorul ogivelor corespunzătoare distribuţiilor considerate. Ogiva prezintă faţă de poligonul de frecvenţă avantajul că se construieşte la fel, indiferent dacă scara distribuţiei este formată din intervale egale sau nu.

TABELUL 86

Distribuţii de salarii săptămânale pentru lucrătorii din industria metalurgică

Salarii în lei	Frecvenţa relativă a lucrătorilor	
	Statul A	Statul B
(1)	(2)	(3)
600 - 650	1	-
650 - 700	7	-
700 - 750	32	2
750 - 800	110	15

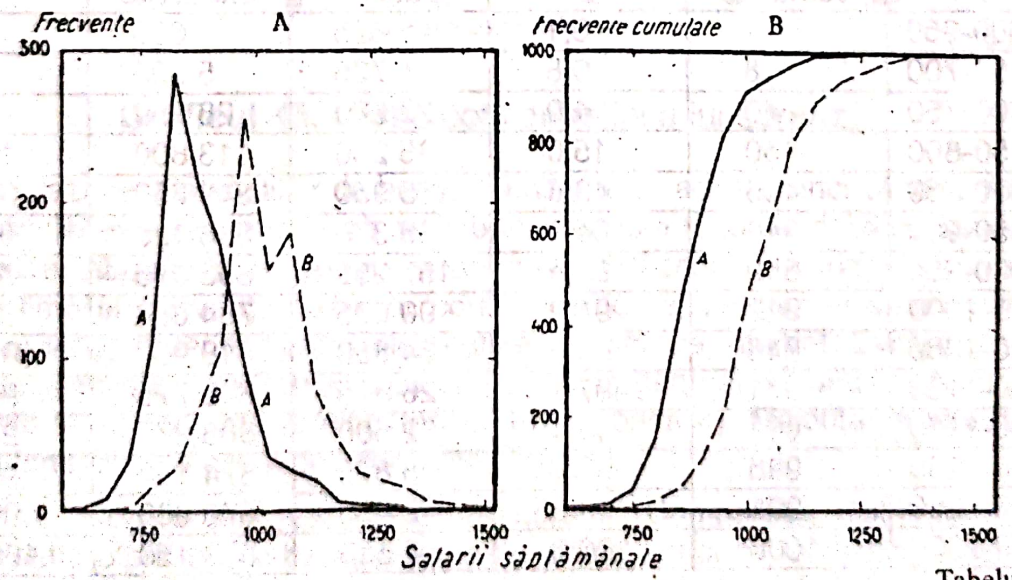
[1] *Comparaţia distribuţiilor în spaţiu este greoaie şi nu se face încă dupa norme generale; despre comparaţia seriilor în timp ne vom ocupa mai târziu (Cap.XIV).*

Scanned with OKEN Scanner

(continuare tab. 86)

(1)	(2)	(3)
800 - 850	286	26
850 - 900	213	66
900 - 950	170	108
950 - 1000	93	256
1000 - 1050	34	154
1050 - 1100	25	181
1100 - 1150	20	82
1150 - 1200	5	46
1200 - 1250	2	23
1250 - 1300	2	18
1300 - 1350	-	14
1350 - 1400	-	5
1400 - 1450	-	3
1450 - 1500	-	1
	1000	1000

Dacă ogiva corespunzătoare unei distribuții se găsește pe grafic mereu sau aproape mereu sub ogiva unei alte distribuții, vom spune că valorile primei serii sunt în general mai mari decât ale celei din urmă. Acesta este cazul distribuțiilor din tabelul 86 (graficul 44 B). Într-adevăr, se găsesc mai mulți lucrători, proporțional vorbind, în statul B, care să primească un salariu mai mic decât o anumită sumă și aceasta, oricare ar fi suma. Este firesc deci să spunem că salariile în A sunt mai mici decât cele din B.



Tabelul 86

Graficul 44. Comparație grafică directă a distribuțiilor heterograde.

Evident că operaţia de comparaţie grafică a seriilor trebuie precedată, în cazul când seriile nu se prezintă cu frecvenţe relative, de calculul acestora din urmă. Comparând frecvenţele relative, eliminăm diferenţele ce ar proveni din faptul că seriile nu cuprind acelaşi număr de elemente.

117. *Curba lui Lorenz.* Un alt mod de comparaţie grafică a distribuţiilor heterograde care ne poate duce la concluzii din cele mai interesante consistă din întrebuinţarea curbei lui LORENZ[1].

Curba lui Lorenz nu este, de fapt, decât o variantă a ogivei. Prin ea se pun în legătură frecvenţele cumulate cu valorile cumulate ale variabilei independente, ambele fiind calculate proporţional cu 100, 1000 etc. Acest mod de a reprezenta grafic o distribuţie este special recomandat pentru cazurile de bunuri, venituri, impozite. Obţinem astfel o imagine grafică a concentrării acestor variabile. Să ne referim la exemplul din tabelul 87. Se calculează, în primul rând, frecvenţele cumulate exprimate în procente faţă de totalul frecvenţelor simple (coloana 3), în al doilea rând, se calculează salariul total al fiecărei clase din distribuţie (col. 4) şi, în fine, salariile cumulate exprimate şi ele în procente faţă de totalul lor (col. 6). O dată aceste valori calculate, pe laturile unui pătrat, ca axe de coordonate, se înseamnă respectiv valorile frecvenţelor cumulate şi ale salariilor cumulate (graficul 45). Curba obţinută unind punctele ale căror coordonate le avem pe axele pătratului este curba concentraţiei sau curba lui Lorenz corespunzătoare distribuţiei din tabelul 87.

TABELUL 87

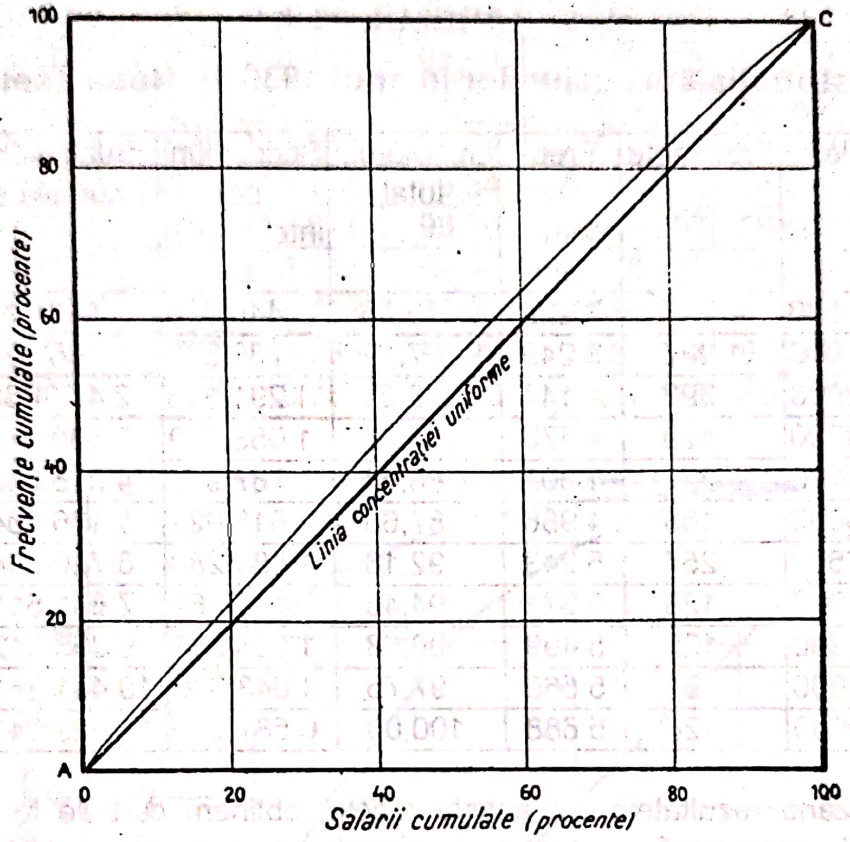
Calculul curbei de concentrare pentru distribuţia salariilor din statul A

Salarii în lei	Frecvenţe cumulate	%	Salariile claselor	Salarii cumulate	%
600-650	1	0,1	625	625	0,1
650-700	8	0,8	4 725	5 350	0,6
700-750	40	4,0	23 200	28 550	3,2
750-800	150	15,0	85 250	113 800	12,9
800-850	436	43,6	235 950	349 750	39,8
850-900	649	64,9	186 375	536 125	61,0
900-950	819	81,9	157 250	693 375	78,9
950-1000	912	91,2	90 675	784 050	89,2
1000-1050	946	94,6	34 850	818 900	93,1
1050-1100	971	97,1	26 875	845 775	96,2
1100-1150	991	99,1	22 500	868 275	98,8
1150-1200	996	99,6	5 875	874 150	99,4
1200-1250	998	99,8	2 450	876 600	99,7
1250-1300	1000	100,0	2 550	879 150	100,0

[1] *M. O. Lorenz,* ***Methods of Measuring the Concentration of Wealth,*** *Quarterly Publication of the American Statistical Association, 1905, p. 209-219.*

Scanned with OKEN Scanner

Diagonala AC corespunde unei distribuţii în care 10% din populaţie posedă 10% din averea totală, 20% din populaţie, 20% din averea totală etc. Din această cauză, AC poartă numele de *linia concentraţiei uniforme.* Depărtarea curbei lui Lorenz faţă de diagonala AC ne arată pentru fiecare caz în parte concentrarea averilor, a veniturilor etc.



Tabelul 87

Graficul 45. Curba concentraţiei a lui Lorenz.

În cazurile distribuţiilor de bunuri, venituri etc., adică atunci când problema concentraţiei este prima care să ne intereseze, este recomandabil ca distribuţia să fie astfel tabulată încât, pentru fiecare interval al scării de variaţie, să avem în acelaşi timp şi frecvenţa cazurilor corespunzătoare, precum şi suma valorilor variabilei independente pentru elementele din fiecare interval. Evităm astfel o evaluare a acesteia din urmă, care se obţine, în general, făcând produsul între frecvenţe şi mijlocul intervalelor (tabelul 87, coloana 4). Majoritatea statisticilor oficiale îndeplineşte acest deziderat metodologic.

118. *Comparaţia prin curba lui Lorenz.* Comparaţia distribuţiilor prin instrumentul oferit de curba lui Lorenz ne poate conduce la rezultate din cele mai interesante în ceea ce priveşte concentraţia repartiţiei diferitelor bunuri. S-a ajuns pe această cale să se stabilească faptul că, în general şi cu foarte puţine excepţii, concentrarea veniturilor este mai aproape de distribuţia uniformă decât

Scanned with OKEN Scanner

aceea a averilor, iar concentrarea acesteia din urmă este mai redusă decât cea înfățișată de cuantumul moștenitorilor[1].

Ca o ilustrare a procedeului menționat, să considerăm curbele lui Lorenz pentru distribuțiile de venituri și moșteniri din Noua Zeelandă, 1930 (tabelele 88 A și 88 B).

TABELUL 88 A

Distribuția succesiunilor în anul 1930, în Noua Zeelandă

Intervale în £	Frecvențe	Frecvențe cumulate	La sută din total	Succesiuni pe intervale în £	Succesiuni cumulate	La sută din total
Sub 500	2 222	2 222	39,06	446 378	446 378	2,62
500-1000	1 027	3 249	57,12	739 372	1 185 750	6,96
1000-2000	898	4 147	72,21	1 291 913	2 477 663	14,55
2000-3000	429	4 576	80,45	1 053 009	3 530 672	20,73
3000-4000	227	4 803	84,44	787 971	4 318 643	25 36
4000-5000	183	4 986	87,66	818 021	5 136 664	30,16
5000-7500	257	5 243	92,18	1 584 280	6 720 944	39,47
7500-10000	128	5 371	94,43	1 110 567	7 831 511	45,99
10000-15000	128	5 499	96,68	1 566 531	9 398 042	55,19
15000-20000	61	5 560	97,75	1 043 319	10 441 361	61,32
Peste 20000	128	5 688	100,00	6 587 680	17 029 041	100,00

Utilizând rezultatele din aceste tabele, obținem curbele lui Lorenz din graficul 46, care confirmă faptul menționat mai sus: concentrația este mai pronunțată pentru succesiuni decât pentru venituri.

Tabelul 88 B

Distribuția venitului total în anul 1930, în Noua Zeelandă[2]

Intervale în £	Frecvențe	Frecvențe cumulate	La sută din total	Venitul total pe intervale	Venituri cumulate	La sută din total
(1)	(2)	(3)	(4)	(5)	(6)	(7)
Sub 300	38 867	38 867	32,73	8 413 044	8 413 044	12,71
300-400	34 883	73 750	62,10	11 938 486	20 351 530	30,73

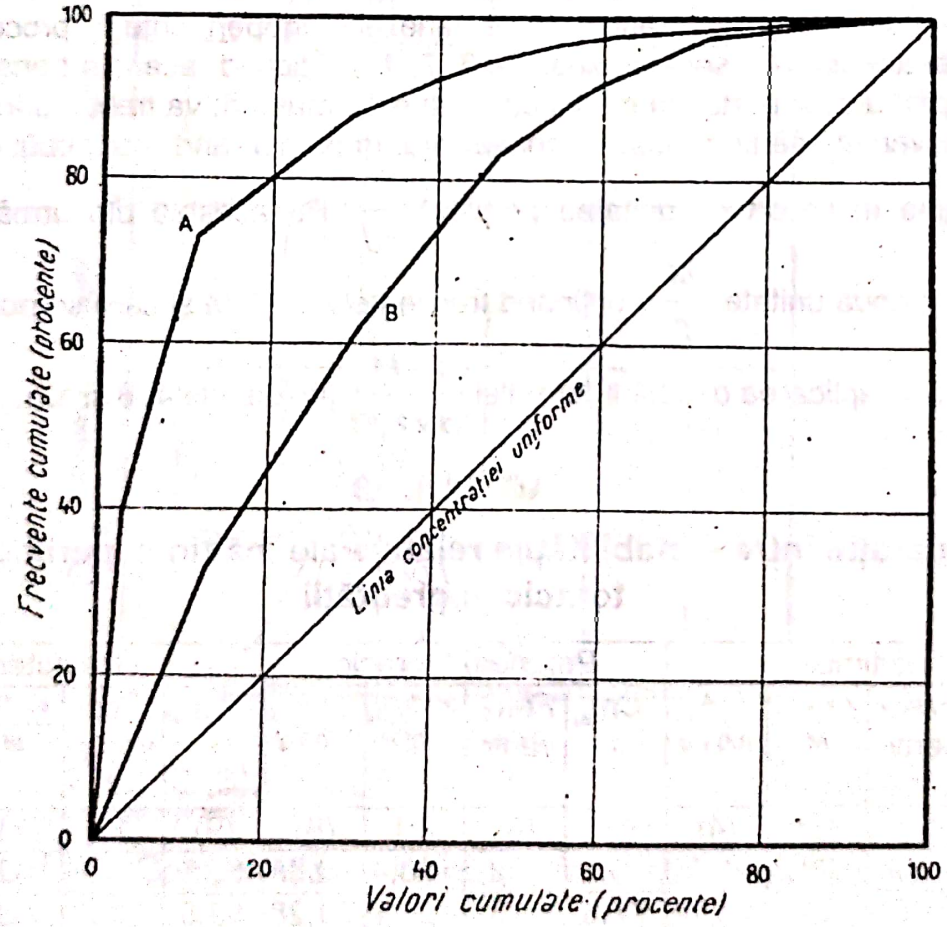
[1] *W. I. King, **Wealth Distribution in the Continental United States at the Class of 1921**, Journal of the American Statistical Association, June, 1927 (XXII) p.135-153.*

[2] *Datele care au servit ca bază calculelor din aceste tabele au fost extrase din **New Zeeland Official Year Book**, 1932, p. 643, 658-659.*

Scanned with OKEN Scanner

(continuare tab. 88 B)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
400-500	16 578	90 328	76,06	7 304 240	27 655 770	41,76
500-600	8 201	98 529	82,96	4 446 020	32 101 790	48,48
600-700	4 967	103 496	87,15	3 196 184	35 297 974	53,31
700-800	3 164	106 660	89,81	2 356 546	37 654 520	56,86
800-900	2 215	108 875	91,68	1 869 438	39 523 958	59,69
900-1000	1 512	110 387	92,95	1 428 567	40 952 525	61,85
1000-2000	5 729	116 179	97,83	7 720 343	48 672 868	73,50
2000-5000	2 036	118 215	99,54	5 843 651	54 516 519	82,33
5000-10000	296	118 511	99,79	1 946 097	56 462 616	85,27
10000-20000	132	118 643	99,90	1 806 735	58 269 351	88,00
20000-50000	70	118 713	99,96	2 088 248	60 357 599	91,15
50000-100000	26	118 739	99,98	1 882 944	62 240 543	93,99
Peste 100000	18	118 757	100,00	3 977 519	66 218 062	100,00



A = concentrația succesiunilor.
B = concentrația veniturilor totale.

Tabelele 88 A și 88 B

Graficul 46. Comparația concentrației prin curba lui Lorenz.

Scanned with OKEN Scanner

119. *Comparația variabilității relative a distribuțiilor heterograde.* O comparație a distribuțiilor heterograde se poate face și din punctul de vedere al dispersiei, dar mai ales din acel al *variabilității relative*. Numeric, variabilitatea relativă se măsoară prin coeficientul de variabilitate (§ 73). Astfel, pentru cazul distribuțiilor înălțimilor, perimetrului toracic și greutății ale unui grup de soldați români (tabelul 89), avem:

Înălțimea..... C. V. = 21,99

Perimetrul toracic .... C. V. = 42,96

Greutatea ..... C. V. = 88,02

ceea ce ne arată că variabilitățile relative ale acestor elemente biometrice sunt complet diferite între ele. Există însă și un procedeu grafic prin care să se pună în evidență deosebirile constatate numeric prin intermediul coeficientului de variabilitate, procedeu care are avantajul de a prezenta lucrurile sub o formă mai intuitivă.

Cum ceea ce interesează din punctul de vedere al variabilității relative este mărimea abaterii în raport cu media respectivă, și nu abaterea absolută, vom transforma în primul rând valorile variabilei independente în procente față de media aritmetică a seriei (coloanele 3, 7, 11). Făcând această transformare, am adoptat a suta parte din medie ca unitate de măsură; va trebui, prin urmare, ca și frecvențele să fie reduse în consecință. În primul rând, vor trebui calculate frecvențele în raport cu unitatea primitivă, $\frac{f}{d}$. Pe acestea din urmă le vom înmulți cu noua unitate $\frac{M}{100}$, obținând frecvențele căutate și care se pot calcula, de fapt, prin aplicarea directă a formulei $\frac{f \times M}{d \times 100}$ (coloanele 4, 8 și 12).

TABELUL 89

Comparație între variabilitățile relative ale înălțimii, perimetrului toracic și greutății

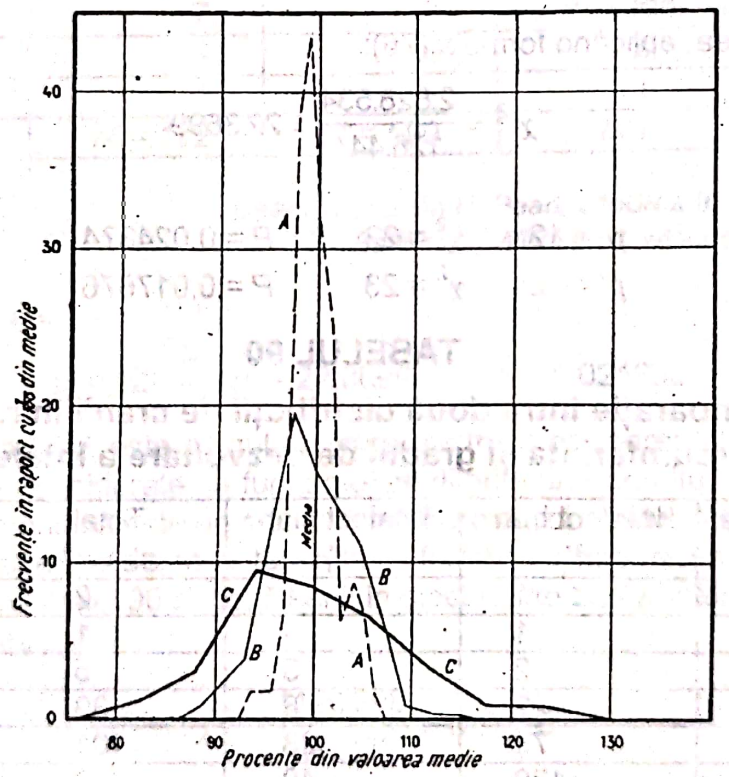
Înălțimea				Perimetrul toracic				Greutatea			
Cm x	Frec-vențe f	$\frac{x \times 100}{M}$	$\frac{f \times M}{100 \times d}$	Cm x	Frec-vențe f	$\frac{x \times 100}{M}$	$\frac{f \times M}{100 \times d}$	Kg x	Frec-vențe f	$\frac{x \times 100}{M}$	$\frac{f \times M}{100 \times d}$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
162,5	2	93,5	1,74	75,5	2	88,4	0,85	55,5	7	82,1	1,18
164,5	2	94,7	1,74	77,5	3	90,7	1,28	59,5	18	88,0	3,04
166,5	2	95,8	1,74	79,5	9	93,1	3,84	63,5	56	93,9	9,46
168,5	8	97,0	6,95	81,5	28	95,4	11,96	67,5	50	99,9	8,45
170,5	44	98,1	38,22	83,5	46	97,7	19,65	71,5	38	105,8	6,45

Scanned with OKEN Scanner

(continuare tab. 89)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
172,5	50	99,3	43,43	85,5	37	100,1	15,80	75,5	20	111,7	3,38
174,5	36	100,4	31,27	87,5	31	102,4	13,24	79,5	5	117,6	0,85
176,5	29	101,6	25,19	89,5	26	104,8	11,11	83,5	5	123,5	0,85
178,5	7	102,7	6,08	91,5	14	107,1	5,98	87,5	1	122,4	0,17
180,5	10	103,9	8,69	93,5	2	109,4	0,85	-	-	-	-
182,5	8	105,0	6,95	95,5	1	111,8	0,43	-	-	-	-
184,5	2	106,2	1,74	97,5	1	111,1	0,43	-	-	-	-
$M = 173,73$ $\sigma = 3,82$				$M = 85,43$ $\sigma = 3,67$				$M = 67,60$ $\sigma = 5,95$			

Reprezentarea grafică a distribuțiilor se face în urmă pe baza noilor frecvențe și a procentelor față de medie (graficul 47). Pe un astfel de grafic, mediile aritmetice sunt reduse la aceeași valoare egală cu 100, iar - un lucru esențial - suprafețele închise de poligoanele de frecvență sunt echivalente[1].



A = Distribuția taliei.
B = Distribuția perimetrului toracic.
C = Distribuția greutății.

Tabelul 89

Graficul 47. Comparația variabilității distribuțiilor heterograde.

[1] *Dacă totalul frecvențelor n-ar fi fost același pentru toate distribuțiile, ar fi trebuit să calculăm în prealabil frecvențele relative, tocmai pentru a obține această echivalență, fără de care comparația pierde orice bază logică.*

Scanned with OKEN Scanner

Metoda aceasta prezintă avantajul de a permite comparaţia distribuţiilor heterograde, chiar dacă acestea sunt făcute în raport cu variabile de structuri diferite.

120. *Comparaţia distribuţiilor de frecvenţă prin intermediul coeficientului de satisfacţie P.* Pentru o comparaţie prin calcul a distribuţiilor de frecvenţă, ne vom servi de acelaşi criteriu pe care l-am întrebuinţat pentru comparaţia frecvenţelor statistice cu cele ajustate (§ 87). Într-adevăr, formal şi chiar în fond, aceste două probleme sunt analoage.

Dacă n'_i şi n''_i sunt frecvenţele observate pentru acelaşi interval în cele două distribuţii, vom calcula în primul rând:

$$\chi^2 = \frac{1}{N'N''} \sum \frac{(n'_i N'' - n''_i N')^2}{n'_i + n''_i} \qquad (76)$$

unde N' şi N'' reprezintă totalul frecvenţelor din cele două distribuţii. Valoarea lui P este dată de tablele menţionate, adoptându-se valoarea p' egală cu numărul intervalelor considerate. Să ne referim, pentru exemplificarea acestui procedeu, la distribuţiile din tabelul 90.

Vom avea, aplicând formula (76):

$$\chi^2 = \frac{2.828.534}{126.444} = 22{,}3699$$

iar din table:

$p' = 12 \quad \chi^2 = 22 \quad P = 0{,}024374$

$p' = 12 \quad \chi^2 = 23 \quad P = 0{,}017676$

TABELUL 90

Comparaţie între două distribuţii de cranii în raport cu circumferinţa şi gradul de dezvoltare a intelectului

Circumferinţa în cm	Intelect bun n'_i	Intelect slab n''_i	Total $n'_i + n''_i$	$\frac{(n'_i N'' - n''_i N')^2}{n'_i + n''_i}$
50-51	1	1	2	184 225
51-52	1	-	1	26 896
52-53	2	3	5	788 045
53-54	22	8	30	218 453
54-55	78	23	101	241 718
55-56	139	40	179	361 486
56-57	236	46	282	37 180
57-58	176	28	204	259 511
58-59	88	13	101	192 468
59-60	25	1	26	426 240
60-61	2	1	3	65 416
61-62	1	-	1	26 896
	$N' = 771$	$N'' = 164$	935	2 828 534

Scanned with OKEN Scanner

TABELUL 91

Comparaţia circumferinţei craniene pentru cazurile de intelect foarte bun şi intelect bun

Circumferinţa în cm	Intelect foarte bun n'_i	Intelect bun n''_i	Total $n'_i + n''_i$	$\frac{(n'_i N'' - n''_i N')^2}{n'_i + n''_i}$
50-51	-	1	1	29.584
51-52	1	-	1	358.801
52-53	-	2	2	59.168
53-54	3	19	22	98.356
54-55	17	61	78	1.224
55-56	27	112	139	68.736
56-57	54	182	236	4.601
57-58	40	136	176	1.833
58-59	22	66	88	37.890
59-60	7	18	25	48.136
60-61	-	2	2	59.168
61-62	1	-	1	358.801
	$N' = 172$	$N'' = 599$	771	1.126.298

Date extrase din R. Pearl, ***Studies in Human Biology,*** *Baltimore, Williams & Wilkins Co., 1924, p.97.*

şi prin interpolare:

$$p' = 12 \qquad \chi^2 = 22{,}3699 \qquad P = 0{,}021896.$$

Valoarea lui P este destul de mică pentru a ne îndreptăţi să privim cele două distribuţii considerate ca fundamental diferite una de alta. Cu alte cuvinte, circumferinţa craniană este un element ce diferenţiază gradul de inteligenţă.

Să considerăm acum şi tabelul 91, format de împărţirea elementelor cu intelect bun din tabelul 90 în două serii: intelect foarte bun şi intelect bun.

Vom avea :

$$\chi^2 = \frac{1.126.298}{103.028} = 10{,}9320$$

şi deci:

$$p' = 12 \qquad \chi^2 = 10{,}9320 \qquad P = 0{,}339187$$

Această valoare a lui P este sensibil diferită de cea obţinută mai sus. Probabilitatea ca cele două distribuţii din tabelul 91 să fie eşantioanele aceleiaşi colectivităţi nu mai poate fi neglijată. Completând rezultatul deja obţinut, vom

Scanned with OKEN Scanner

spune că mărimea circumferinței craniene diferențiază intelecte bune și foarte bune de cele slabe, nu însă și pe cele foarte bune de cele bune.

În ceea ce privește acum o limită sub care valorile lui P să fie privite ca semnificative, în sensul deosebirii fundamentale a două distribuții de frecvență, se adoptă în general valoarea 0,05. Vedem dar că și aici, ca și pentru abateri (§ 85), s-a ales ca limită o probabilitate egală aproximativ cu $\frac{1}{20}$.

121. *Cazul distribuțiilor homograde.* Procedeul expus în paragraful precedent necesitând numai cunoașterea frecvențelor claselor respective, se va putea aplica întocmai și distribuțiilor homograde.

Să considerăm tabelul 92. Calculul lui χ^2 se va face după formula (76), pornind de la frecvențele celor două distribuții.

TABELUL 92

Distribuția sexelor în raport cu culoarea părului

(Copii scoțieni)

Culoarea părului	Băieți n'_i	Fete n''_i	Total $n'_i + n''_i$	$\frac{(n'_i N'' - n''_i N')^2}{n'_i + n''_i}$
Blond	592	544	1 136	6 642 037
Roșcat	119	97	216	332 683
Castaniu	849	677	1 526	5 554 608
Închis	504	451	955	2 459 840
Negru	36	14	50	24 204 099
	2 100	1 783	3 883	39 193 267

Datele reproduse după R. A. Fisher, ***op. cit.,*** *p. 90.*

De unde:

$$p' = 5 \quad \chi^2 = 10{,}47 \qquad P = 0{,}033931.$$

Deducem, prin urmare, că sexul influențează categoric culoarea părului, cel puțin pentru categoria examinată.

122. *Principiul comparației valorilor statistice.* Vom trece acum la metodele de comparație a valorilor caracteristice. Din cauza importanței covârșitoare pe care o au în statistica matematică, ne vom ocupa în special de cazul frecvențelor relative și al mediilor aritmetice.

Problema generală se pune în felul următor:

Urmărind manifestările unui fenomen colectiv în două circumstanțe deosebite, obținem două serii statistice cu ajutorul cărora calculăm două valori caracteristice c și c'. *Diferența $c - c'$ este datorată fluctuațiilor hazardului sau, dimpotrivă, este semnificativă, în sensul că evidențiază o diferență de fond între manifestările fenomenului considerat, diferență care este consecința circumstanțelor diferite în care s-au produs acestea?*

Scanned with OKEN Scanner

Soluția acestei probleme se bazează pe proprietatea că dacă σ_c și $\sigma_{c'}$ sunt abaterile-tip ale valorilor c, c', abaterea-tip a diferenței $c - c'$ va fi dată de formula:

$$\sigma^2_{c-c'} = \sigma^2_c + \sigma^2_{c'} \tag{77}$$

cu condiția esențială ca cele două eșantioane *să fie obținute la întâmplare și independent unul de altul.*

Cunoscând abaterea tip a diferenței $c - c'$, vom aplica acesteia proprietatea variației statistice (§ 85), adică vom considera diferența $c - c'$ ca semnificativă, dacă:

$$|c - c'| \geq 2\sigma_{c-c'} \tag{78}$$

În cazul contrar, nu vom putea face nici o afirmație în privința acestei diferențe. Aceasta poate fi datorată în acest caz tot așa de bine diferenței de circumstanțe a manifestărilor fenomenului considerat, ca și fluctuațiilor hazardului. Vom vedea mai târziu care este mijlocul de a evita suprapunerea acestor două posibilități.

123. *Comparația frecvențelor statistice.* Să examinăm cazul comparației frecvențelor relative, f_1, f_2, ale aceluiași eveniment în două grupuri distincte. Dacă frecvențele teoretice corespunzătoare acestora sunt p_1, p_2, vom avea (57):

$$\sigma_1 = \sqrt{\frac{p_1 q_1}{n_1}} \qquad \sigma_2 = \sqrt{\frac{p_2 q_2}{n_2}}$$

unde n_1, n_2 reprezintă numărul cazurilor din fiecare grup; deci abaterea tip a diferenței $f_1 - f_2$ va fi (77):

$$\sigma_{12} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}. \tag{79}$$

Aplicarea acestei formule întâmpină dificultăți prin faptul că noi ignorăm în marea majoritate a cazurilor valorile p_1, p_2. O evaluare a lui σ_{12} se poate face în două moduri:

a) p_1, p_2 sunt înlocuiți prin valorile statistice f_1, f_2 și, prin urmare:

$$\sigma_{12} = \sqrt{\frac{f_1(1-f_1)}{n_1} + \frac{f_2(1-f_2)}{n_2}}. \tag{79 bis}$$

b) Presupunând că frecvența teoretică a evenimentului considerat este aceeași în ambele grupuri, vom avea:

Scanned with OKEN Scanner

$$\sigma_1 = \sqrt{\frac{p_0 q_0}{n_1}} \qquad \sigma_2 = \sqrt{\frac{p_0 q_0}{n_2}}$$

și deci:

$$\sigma_{12} = \sqrt{f_0(1-f_0)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \qquad \text{(79 ter)}$$

unde:

$$f_0 = \frac{n_1 f_1 + n_2 f_2}{n_1 + n_2}$$

valoare care reprezintă frecvența rezultând din compunerea celor două grupuri.

Trecând acum la o aplicație numerică a acestui procedeu, să considerăm tabelul 93.

TABELUL 93

Legătura între inoculare contra holerei și contagiune

	S-au contaminat	Nu s-au contaminat	Total
Inoculați	276	3	279
Neinoculați	473	66	539
Total	749	69	818

Reprodus după U.G. Yule, ***op. cit.****, p. 31.*

Vom avea frecvențele relative ale celor contaminați în cele două grupuri, inoculați și neinoculați:

$$f_1 = \frac{276}{279} = 0{,}989 \qquad f_2 = \frac{473}{539} = 0{,}878$$

iar:

$$f_0 = \frac{749}{818} = 0{,}916.$$

Deci, după (79 bis):

$$\sigma_{12} = \sqrt{\frac{0{,}989 \times 0{,}011}{279} + \frac{0{,}878 \times 0{,}122}{539}} = 0{,}015$$

Scanned with OKEN Scanner

sau după (79 ter):

$$\sigma_{12} = \sqrt{0{,}916 \times 0{,}084 \times \left(\frac{1}{279} + \frac{1}{539}\right)} = 0{,}020.$$

Prima formulă este bazată pe supoziţia *a priori* că diferenţa $f_1 - f_2$ este semnificativă; avem:

$$f_1 - f_2 = 0{,}111 = 7{,}4 \times 0{,}015$$

ceea ce confirmă ipoteza făcută.

Dacă presupunem că diferenţa între f_1 şi f_2 nu este semnificativă, ci datorată hazardului, atunci vom întrebuinţa a doua formulă; or:

$$f_1 - f_2 = 0{,}111 = 5{,}55 \times 0{,}02$$

de unde se deduce că nu putem considera pe f_1 şi f_2 ca evaluările statistice ale aceleiaşi frecvenţe teoretice p_0. Prin urmare, influenţa inoculării contra holerei asupra contaminării în caz de epidemie este *statistic* stabilită. Să considerăm datele din tabelul 94. Vom avea:

$$f_1 = \frac{47}{68} = 0{,}691 \; f_2 \frac{12}{15} = 0{,}800 \; f_0 = \frac{59}{83} = 0{,}711$$

şi deci:

$$\sigma_{12} = 0{,}141 \quad (79 \text{ bis}) \qquad \sigma_{12} = 0{,}129 \quad (79 \text{ ter}).$$

În ambele ipoteze, diferenţa între f_1 şi f_2 nu este semnificativă:

$$f_1 - f_2 = 0{,}109 = 0{,}77 \times 0{,}141 \; (79 \text{ bis})$$
$$= 0{,}84 \times 0{,}129 \; (79 \text{ ter})$$

Vom conclude, prin urmare, că diferenţa $f_1 - f_2$ este datorată fie hazardului, fie unei diferenţe de fond, care însă, dacă există, este mai mică decât intervalul de variaţie statistică permisă în cazul exemplului citat.

TABELUL 94

Legătura între culoarea florii şi caracterul suprafeţei fructului plantei *Datura*

Suprafaţa fructului	Culoarea florii		Total
	violetă	albă	
Aspră	47	21	68
Netedă	12	3	15
Total	59	24	83

Reprodus din ***Tables for Statisticians and Biometricians****, Part I, p. XXXIV.*

Scanned with OKEN Scanner

Observăm că σ_{12}, calculat prin ambele formule de mai sus, descreşte când n_1 şi n_2 cresc, adică atunci când eşantioanele devin mai numeroase. Mărind, prin urmare, numărul cazurilor observate, vom obţine o valoare oricât de mică dorim pentru σ_{12}, astfel încât, dacă diferenţa $p_1 - p_2$ este reală, ajungem la un moment când $2\sigma_{12} < |p_1 - p_2|$. În acest caz, intervalul de variaţie statistică a diferenţei $f_1 - f_2$ nu va mai putea ascunde diferenţa esenţială dintre frecvenţele considerate, dacă această diferenţă există într-adevăr. Iată deci care este mijlocul de a înlătura îndoiala în privinţa unei diferenţe nesemnificative: *mărirea eşantionului*. Din nenorocire, acest procedeu nu se poate aplica tuturor cazurilor, căci experimentarea asupra fenomenelor colective nu este întotdeauna posibilă. Îndoiala va rămâne atunci şi nu va putea fi înlăturată decât în parte şi pe baza considerentelor de altă natură.

124. *Comparaţia mediilor aritmetice.* Problema care ne preocupă se pune în acelaşi mod şi pentru cazul mediilor aritmetice. Dacă σ_1, σ_2 sunt abaterile tip ale populaţiilor din care am extras două eşantioane ale căror medii aritmetice şi abateri-tip sunt respectiv m_1, m_2; s_1, s_2, abaterea tip a diferenţei $m_1 - m_2$ va fi (77, 61):

$$\sigma_{12} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (80)$$

care, din cauză că σ_1, σ_2 nu pot fi cunoscuţi întotdeauna, este înlocuită prin:

$$\sigma_{12} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad (80 \text{ bis})$$

Să considerăm distribuţiile din tabelul 90. Vom avea:

pentru clasa intelect bun: $m_1 = 56{,}578$ $\quad \frac{s_1^2}{n_1} = 0{,}002495$

pentru clasa intelect slab: $m_2 = 56{,}091$ $\quad \frac{s_2^2}{n_2} = 0{,}014833$ [1]

şi deci:

$$\sigma_{12} = \sqrt{0{,}017328} = 0{,}132$$

cum:

$$m_1 - m_2 = 0{,}487 = 3{,}7 \times 0{,}132$$

[1] *Fluctuaţiile s^2 au fost calculate ţinându-se seamă de corecţia Sheppard.*

Scanned with OKEN Scanner

deducem că această diferenţă este semnificativă. Se confirmă astfel rezultatul obţinut pe altă cale, mai sus (§ 120).

Considerând acum distribuţiile din tabelul 91, obţinem:

pentru clasa intelect foarte bun $m_1 = 56{,}703$ $\frac{s_1^2}{n_1} = 0{,}011546$

pentru clasa intelect bun $m_2 = 56{,}542$ $\frac{s_2^2}{n_2} = 0{,}003398$

de unde:

$$\sigma_{12} = \sqrt{0{,}014944} = 0{,}122$$

$$m_1 - m_2 = 0{,}161 = 1{,}3 \times 0{,}122$$

fapt care conduce la concluzia că diferenţa circumferinţei craniene nu este semnificativă în această din urmă privinţă.

Diferenţa $m_1 - m_2$ poate proveni fie din cauza hazardului, fie dintr-o diferenţă reală, care însă este de un ordin mai mic decât acel al diferenţei ce poate fi plauzibil atribuită hazardului. Acest echivoc se poate înlătura prin mijlocul pomenit cu ocazia examinării aceleiaşi chestiuni în cazul diferenţelor dintre frecvenţele relative, adică prin mărirea numărului cazurilor examinate.

Într-adevăr, dacă diferenţa între două medii m_1, m_2 are la bază o diferenţă reală între mediile ideale corespunzătoare, atunci se poate obţine, mărind pe n_1, n_2, o valoare a lui σ_{12} (80) cu mult inferioară acestei diferenţe. Dacă pe măsură ce mărim numărul observaţiilor se micşorează diferenţa $m_1 - m_2$, atunci înseamnă că această diferenţă este datorată fluctuaţiilor permise produse de hazard; dacă, dimpotrivă, această diferenţă este o consecinţă a diferenţei mediilor ideale, atunci va sosi un moment în care, mărind mereu numărul observaţiilor, vom obţine pentru σ_{12} o valoare suficient de mică pentru ca diferenţa $m_1 - m_2$ să fie mai mare decât $2\sigma_{12}$.

125. *Comparaţia statistică a seriilor puţin numeroase.* Vom trece acum să examinăm un ultim caz de comparaţie a seriilor statistice. Acest gen de probleme se întâlneşte mai ales în terapeutică şi psihologia aplicată. Îl vom ilustra printr-un exemplu.

S-au administrat la 10 pacienţi doze de dextro- şi levo-hyoscyamină hydrobromidă în scopul de a le produce un somn mai lung. Orele de somn suplimentare produse prin administrarea acestor două medicamente sunt înfăţişate în tabelul 95[1].

[1] *Rezultatele sunt obţinute de A. R. Cushney şi A. R. Peebles (**Journal of Physiology**, 1904). Noi le reproducem din **Tables for Statisticians and Biometricians**, Part II, p. CXXXVI.*

Scanned with OKEN Scanner

Problema care se pune în legătură cu aceste rezultate se compune din două părți. Trebuie stabilit, în primul rând, dacă medicamentele întrebuințate sunt într-adevăr eficace și în al doilea, care din acestea este mai indicat scopului urmărit.

Să considerăm cazul pentru dextro-hyoscyaminã. Din cauză că numărul cazurilor observate este relativ mic, 10, nu vom mai putea aplica regula variației statistice a mediei, așa cum a fost expusă mai sus (§ 102) și care, după cum am precizat la timp, nu este valabilă decât în cazul eșantioanelor numeroase. În cazul eșantioanelor mici, media eșantionului nu se mai distribuie normal. Au fost însă calculate table care să ne permită și în acest caz un studiu al diferenței între media statistică și media teoretică (*Tables for Statisticians and Biometricians*, Part I, p. 36). Procedeul de urmat este următorul: să presupunem că administrarea de dextro-hyoscyaminã este neutră în ceea ce privește provocarea de somn. Atunci media teoretică a distribuției (1) din tabelul 95 va fi nulă și deci diferența acesteia față de media statistică va fi chiar 0,75.

TABELUL 95

Orele suplimentare de somn obținute prin administrarea de hyoscyaminã hydrobromidă

Pacient	Dextro- (1)	Levo- (2)	Diferența (2)-(1)
1	+0,7	+1,9	+1,2
2	- 1,6	+0,8	+2,4
3	- 0,2	+1,1	+1,3
4	- 1,2	+0,1	+1,3
5	- 0,1	- 0,1	0,0
6	+3,4	+4,4	+1,0
7	+3,7	+5,5	+1,8
8	+0,8	+1,6	+0,8
9	0,0	+4,6	+4,6
10	+2,0	+3,4	+1,4
Media	+0,75	+2,33	+1,58
Fluctuația	2,8805	3,6081	1,3616
Abaterea-tip	1,70	1,90	1,17

Se calculează raportul între această diferență și abaterea-tip a eșantionului:

$$z = \frac{0,75}{1,70} = 0,44.$$

Din tablele menționate, pentru numărul de cazuri examinate $n = 10$ și pentru această valoare a lui z, obținem posibilitatea ca diferența mediei statistice

față de media teoretică să fie mai mică decât de 0,44 de ori abaterea tip a eșantionului. Avem:

$$n = 10 \qquad z = 0{,}4 \qquad P = 0{,}86961$$

$$n = 10 \qquad z = 0{,}5 \qquad P = 0{,}91607$$

și, prin interpolare liniară:

$$n = 10 \qquad z = 0{,}44 \qquad P = 0{,}88859$$

ceea ce înseamnă că dacă ipoteza făcută este adevărată - media teoretică egală cu 0 - numai în aproximativ 11 cazuri din 100 nu vom avea o valoare mai mare pentru media eșantionului. Aceasta ne îndreptățește să afirmăm că media 0,75 nu este o evaluare a unei medii teoretice nule, adică medicamentul considerat este eficace. Nu vom putea face însă această afirmație cu cvasicertitudine, știut fiind că cvasicertitudinea este reprezentată printr-o probabilitate cel puțin egală cu 0,95 (§ 85).

Același calcul făcut pentru cazul levo-hyoscyaminei hydrobromide ne dă rezultatul:

$$n = 10 \qquad z = 1{,}23 \qquad P = 0{,}99745$$

care nu poate fi interpretat decât în sensul că medicamentul este cert eficace.

Ne mai rămâne, în sfârșit, să comparăm eficacitatea celor două medicamente.

Pentru aceasta am fi poate tentați să aplicăm procedeul din paragraful precedent. Am comite astfel o mare eroare. Într-adevăr, experiențele considerate s-au făcut pe *aceiași pacienți și, prin urmare, rezultatele obținute în cele două cazuri nu mai pot fi privite ca independente, condiție esențială pentru ca formula (77) să fie valabilă.* Ne vom fixa atunci atenția asupra seriei din ultima coloană, care a putut fi obținută pentru motivul că aceiași pacienți au servit ca obiect și primei și celei de a doua experiențe[1]. Dacă ambele medicamente ar fi egal de eficace, media teoretică a diferenței (2) - (1) va fi nulă. Repetând raționamentul de mai sus, vom calcula cu ajutorul tablelor menționate:

$$n = 10 \qquad z = \frac{1{,}58}{1{,}17} = 1{,}35 \qquad P = 0{,}99852\,.$$

Levo-hyoscyamina este, prin urmare, pozitiv mai eficace decât dextro-hyoscyamina.

---

[1] *De aici se desprinde necesitatea ca datele rezultând din experiențe de acest gen să fie prezentate* în mod individual în raport cu fiecare caz examinat. *În caz contrar, concluzia cea mai importantă la care aceste date pot servi nu se mai poate desprinde din rezultate.*

Scanned with OKEN Scanner

126. *Cazul seriilor cuprinzând mai mult de 10 observaţii.* Tablele care ne dau valorile lui P n-au fost calculate decât pentru valorile lui $n \leq 10$. Pentru valorile lui $n > 10$, se obţine P cu o bună aproximaţie prin formula:

$$P = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z'} e^{-\frac{x^2}{2}} dx \qquad (81)$$

unde:

$$z' = z\sqrt{n-3}.$$

Valoarea lui P se obţine cu ajutorul tablelor de la sfârşitul cărţii de faţă (Tablele II):

$$P = \frac{1}{2} + S_1(z').$$

Astfel, în cazul diferenţei (2) - (1) de mai sus, avem:

$$z' = 1{,}35 \times \sqrt{7} = 3{,}57$$

şi, prin urmare, $P = 0{,}99982$, valoare care nu schimbă întru nimic concluziile deduse din considerarea valorii exacte.

Scanned with OKEN Scanner

PARTEA a III-a

Scanned with OKEN Scanner

Capitolul X

Asociație, contingență

127. *Generalități.* Legile naturale îmbracă întotdeauna forma enunțării unei legături fie între manifestările a două sau mai multor fenomene, fie între prezența și absența anumitor caractere. Spunem, de exemplu, că electricitățile de sens contrar se atrag, iar cele de același sens se resping, că temperatura descreşte cu altitudinea, că imbecilii sunt mai numeroşi printre surdomuți, că într-un corp sănătos avem și o minte sănătoasă etc. Acestea sunt legi cunoscute. Prima este o lege rigidă, se verifică, prin urmare, aplicată oricărui caz. Celelalte sunt legi *stocastice*, ele nu stabilesc decât posibilitatea mai mare a unui caracter de a se uni cu un altul decât cu contrariul acestuia din urmă. Înseamnă, aşadar, că este mai natural să întâlnim într-un corp sănătos o minte sănătoasă, însă contrariul nu trebuie să ne surprindă.

Forma sub care am enunțat legile stocastice de mai sus constată numai existența legăturii dintre fenomene și prin aceasta suntem privați de cunoaşterea unui element destul de interesant în această direcție: *gradul de asociație între cele două fenomene sau caractere*. Prezentul capitol este consacrat mijloacelor de care dispune statistica matematică pentru măsura acestui nou element.

Este clar că o definiție perfect circumscrisă a gradului de legătură între două variabile nu poate fi atinsă. Din această cauză, diferite puncte de vedere au putut sta alături pe această chestiune, fără ca să se decidă care este cel mai adecvat. Și natural, fiecare din ele a condus la un mijloc diferit de măsură a acestei legături.

128. *Coeficientul de asociație.* Ne-am ocupat deja (§ 123) de problema stabilirii dacă diferența frecvențelor aceluiași fenomen în două clase diferite poate să fie considerată ca provenind din întâmplare sau nu. Am văzut care este procedeul care ne conduce la rezolvarea acestei chestiuni. Acest procedeu nu ne permite însă altceva decât să stabilim dacă cele două caractere sunt asociate sau nu. Vom arăta acum mijlocul prin care putem măsura și gradul acestei asociații.

Să considerăm o populație oarecare și să avem în vedere două caractere ale indivizilor ce o alcătuiesc. Să însemnăm cu A, B cele două atribute și cu α, β atributele contrare. Vom distinge printre indivizii populației considerate patru grupe:

1. Elementele care prezintă ambele caractere A și B;
2. Elementele care prezintă caracterul A, fără a-l prezenta și pe B, adică elemente ce prezintă caracterele A și β;

Scanned with OKEN Scanner

3. Elemente ce prezintă caracterele α şi B;
4. Elemente ce nu prezintă nici caracterul A, nici caracterul B, adică prezintă pe α şi β;

Vom nota cu (A) numărul elementelor ce prezintă caracterul A, cu $(A\beta)$ numărul celor ce prezintă pe A şi nu pe B etc. şi cu N numărul total al elementelor observate. Tabelul următor ne arată forma generală sub care se va prezenta o astfel de clasificare pe care o vom numi clasificare în 2×2 clase[1].

Al doilea atribut	Primul atribut: A	Primul atribut: α	Total
B	(AB)	$(B\alpha)$	(B)
β	$(A\beta)$	$(\alpha\beta)$	(β)
Total	(A)	(α)	N

Din acest tabel se desprind următoarele relaţii evidente, de altfel:

$$(AB) + (B\alpha) = (B) \qquad (A\beta) + (\alpha\beta) = (\beta)$$
$$(AB) + (A\beta) = (A) \qquad (B\alpha) + (\alpha\beta) = (\alpha)$$
$$(A) + (\alpha) = N \qquad (B) + (\beta) = N \tag{82}$$

Să presupunem acum că între caracterele considerate nu există nici o asociaţie; atunci frecvenţa relativă a caracterului A, de pildă, va fi aceeaşi, fie că vom considera clasa elementelor ce reprezintă caracterul B sau β, fie că vom considera ansamblul întreg al populaţiei, adică:

$$\frac{(AB)}{(B)} = \frac{(A\beta)}{(\beta)} = \frac{(A)}{N}. \tag{83}$$

Cel mai mic grad de asociaţie între cele două caractere va face ca egalităţile de mai sus să nu mai fie verificate. Diferenţa:

$$\delta = (AB) - \frac{(A)\cdot(B)}{N} \tag{84}$$

va fi egală cu 0 în cazul lipsei complete a asociaţiei şi diferită de 0 în caz contrar. Ţinând seamă de relaţiile (82), egalitatea (84) se mai poate scrie:

$$\delta = \frac{1}{N}\{[(AB)+(A\beta)+(B\alpha)+(\alpha\beta)]\cdot(AB) - [(AB)+(A\beta)]\times$$
$$\times[(AB)+(B\alpha)]\} = \frac{1}{N}\cdot[(AB)(\alpha\beta) - (A\beta)(B\alpha)] \tag{85}$$

[1] 2×2 fold classification, *în engleză*.

Evident că putem măsura asociația caracterelor din tabelul considerat pornind de la altă diferență, cum ar fi:

$$(A\beta) - \frac{(A)(\beta)}{N}.$$

Un calcul din cele mai simple ne arată că oricare ar fi diferența de la care am porni, obținem la sfârșit aceeași valoare ca mai sus, abstracție făcând de semn. Permanența valorii (85) ne îndreptățește, mai ales dacă ținem seama și de ideea care ne-a condus la ea, să o considerăm ca o măsură a asociației dintre cele două caractere.

Această măsură este însă o măsură absolută, influențată, prin urmare, de mărimea eșantionului considerat. Pentru a obține o măsură relativă a gradului de asociație, YULE propune următorul *coeficient de asociație:*

$$Q = \frac{N\delta}{(AB)(\alpha\beta) + (A\beta)(B\alpha)} = \frac{(AB)(\alpha\beta) - (A\beta)(B\alpha)}{(AB)(\alpha\beta) + (A\beta)(B\alpha)}. \qquad (86)$$

Acest coeficient este, în primul rând, egal cu 0 când $\delta = 0$ și, prin urmare, conservă principiul măsurării asociației prin diferența δ. În al doilea rând, are o scară de variație fixă, cuprinsă între +1 și -1, limite care nu sunt atinse decât în cazul *asociației unilateral perfecte*, când unul din numerele (AB), $(A\beta)$, $(B\alpha)$, $(\alpha\beta)$ este egal cu 0. Acesta este cazul tabelului:

Al doilea atribut	Primul atribut: A	α	Total
B	(AB)	$(B\alpha)$	(B)
β	0	$(\alpha\beta)$	(β)
Total	(A)	(α)	N

$(AB) = (A)$
$(\alpha\beta) = (\beta)$

în care *toate elementele A sunt și elemente B* și când Q = +1. *Cu toate acestea, asociația nu poate fi considerată ca perfectă în adevăratul înțeles al cuvântului, deoarece nu și toate elementele B sunt elemente A.* Acest lucru s-ar întâmpla dacă clasificarea ar da naștere la tabelul:

Al doilea atribut	Primul atribut: A	α	Total
B	(AB)	0	(B)
β	0	$(\alpha\beta)$	(β)
Total	(A)	(α)	N

$(AB) = (A) = (B)$
$(\alpha\beta) = (\alpha) = (\beta)$

Faptul că Q ia valorile extreme fără ca acest lucru să fie îndeplinit face ca acest coeficient să nu aibă toate opiniile alături de el. Într-adevăr, lăsând la o

Scanned with OKEN Scanner

parte acest neajuns care nu se face simţit din cauză că exemplele de asociaţie perfectă sunt aproape inexistente în studiul fenomenelor colective, trebuie să menţionăm că, în special pentru cazurile cu o asociaţie slabă, *Q* acuză valori prea ridicate în raport cu ceilalţi coeficienţi de măsură a legăturii între caractere.

Vom mai adăuga faptul că valoarea lui *Q* rămâne aceeaşi, fie că se porneşte de la frecvenţele relative, fie de la cele absolute. Cu alte cuvinte, *Q* este independent de mărimea eşantionului. Acest fapt constituie un alt dezavantaj al lui Q. Într-adevăr, noi nu putem privi la fel de semnificativă o valoare a lui *Q* egală, de pildă, cu 0,9, obţinută pe baza unui eşantion cuprinzând 10.000 de elemente, cu o alta egală cu prima, însă obţinută în urma a 100 de observaţii numai. Ultima ar putea fi foarte uşor influenţată de fluctuaţiile hazardului, destul de pronunţate pentru un eşantion redus.

În ceea ce priveşte semnul lui *Q*, trebuie să spunem că dacă $Q > 0$, înseamnă că între caracterele *A* şi *B* există o atracţie; aceste caractere vor fugi unul de altul, dacă $Q < 0$. Bineînţeles că această distincţie n-are nici o importanţă, deoarece în tabelul de asociaţie noi putem schimba coloanele şi liniile între ele, ceea ce aduce cu sine o schimbare de semn. Este totuna a spune că surdomuţenia este pozitiv asociată cu imbecilitatea cu a afirma că acelaşi caracter este negativ asociat cu o normală dezvoltare a intelectului

129. *Interpretarea valorilor coeficientului de asociaţie.* Pentru o exemplificare practică a coeficientului de asociaţie, precum şi pentru a ne da seama de felul cum lucrează acest instrument, vom considera tabelele următoare:

TABELUL 96

Suprafaţa fructului	Culoarea florii		
	Violetă	Albă	Total
Aspră	47	21	68
Netedă	12	3	15
Total	59	24	83

TABELUL 97

Culoarea ochilor fiului	Culoarea ochilor tatălui		
	Deschisă	Închisă	Total
Deschisă	471	148	619
Închisă	151	230	381
Total	622	378	1 000

TABELUL 98

Mediul	Case ridicate (în mii)		
	Construite	În construcţie	Total
Urban	4 960	50	5 010
Rural	1 749	12	1 761
Total	6 709	62	6 771

Scanned with OKEN Scanner

TABELUL 99

	Imbecili	Neimbecili	Total
Surdomuţi	451	14795	15246
Nesurdomuţi	48431	32464323	32512754
Total	48882	32479118	32528000

TABELUL 100

		Dezvoltarea fizică		
		Cu defecte	Fără defecte	Total
Intelectul	Slab	888	1186	2074
	Normal	1420	22793	24213
	Total	2308	23979	26287

Reproduse din G. U. Yule, ***op. cit.*** *p. 37, 33, 61, 32, 45.*

Vom avea, prin aplicarea strictă a formulei (86):

Tabelul 96 ..... $Q = -0{,}28$

Tabelul 97 ..... $Q = 0{,}66$

Tabelul 98 ..... $Q = -0{,}19$

Tabelul 99 ..... $Q = 0{,}91$

Tabelul 100 ..... $Q = 0{,}85$

Ţinând seama de faptul că valorile lui *Q* mai mici în valoare absolută decât 0,3 nu ne îndreptăţesc în general la stabilirea unei legături între cele două caractere, vom lăsa deoparte, ca neasociate, caracterele din tabelele 96 şi 98. Valorile lui *Q* cuprinse între 0,3 şi 0,5 ne indică o tendinţă de asociaţie, cele cuprinse între 0,6 şi 0,7 - o asociaţie sigură, cele între 0,7 şi 0,9 - o asociaţie pronunţată, iar cele mai mari decât 0,9 - o asociaţie aproape rigidă. Prin urmare, între culoarea ochilor tatălui şi culoarea ochilor fiului există o asociaţie sigură, între dezvoltarea fizică şi cea intelectuală - o asociaţie pronunţată, iar între surdomutism şi imbecilitate - o asociaţie perfectă.

Scara valorilor lui *Q*, aşa cum am fixat-o mai sus, este cu totul arbitrară şi diferă de la un autor la altul. De fapt, coeficientul de asociaţie *Q* este unul din mijloacele mai rudimentare de măsură a asociaţiei, aşa încât nu poate servi singur de bază unei concluzii în această direcţie. Cu toate acestea, cu timpul, în urma unei utilizări continue a acestui coeficient, fiecare din noi îşi poate forma o intuiţie justă în ceea ce priveşte semnificaţia diferitelor valori ale lui *Q*.

130. *Coeficientul de asociaţie şi modul clasificării.* Una din principalele obiecţii aduse acestui procedeu de măsurare a gradului de asociaţie este faptul că *valoarea lui Q variază după modul cum a fost făcută clasificarea.* Să verificăm acest lucru printr-un exemplu numeric şi în acest scop să considerăm tabelul 101.

Scanned with OKEN Scanner

TABELUL 101

Asociația între culoarea părului la frați

Al 2-lea frate \ Primul frate	Roșu	Blond	Castaniu	Închis	Negru	Total
Roșu	30,5	32	16	12	-	81,5
Blond	23	416	158	67,75	0,25	665
Castaniu	16	158	394	98,25	8,25	674,5
Închis	12	67,75	98,25	328,5	19	525,5
Negru	-	0,25	8,25	19	10	37,5
Total	81,5	665	674,5	525,5	37,5	1 984

Reprodus din K. Pearson, ***On the Theory of Contingency and its Relation to Association and Normal Correlation,*** *Drapers' Company Research Memoirs, Biometric Series I, p. 32.*

TABELUL 101 A

Al 2-lea frate \ Primul frate	Blond	Închis	Total
Blond	416	249	665
Închis	249	1 070	1 319
Total	665	1 319	1 984

Să presupunem că, în loc de această clasificare amănunțită, am fi avut o clasificare a cazurilor examinate în două categorii: păr blond și păr închis (inclusiv roșu). Am fi obținut clasificarea din tabelul 101 A, pentru care $Q = 0{,}755$.

Dacă însă clasificarea s-ar fi făcut în raport cu părul negru și cel de o altă culoare decât acesta, am fi obținut tabelul 101 B, pentru care $Q = 0{,}924$.

TABELUL 101 B

Al 2-lea frate \ Primul frate	Negru	Altă culoare	Total
Negru	10	27,5	37,5
Altă culoare	27,5	1 919	1 946,5
Total	37,5	1 946,5	1 984

Diferența între valorile lui Q, obținute pornind de la aceeași problemă și același material, ne arată că acest procedeu trebuie mânuit cu multă prudență.

131. *Gradul de asociație măsurat prin coeficientul de satisfacție P.* Un alt mijloc de a măsura gradul de asociație între două variabile se bazează pe

acelaşi principiu care ne-a condus la stabilirea unui coeficient de satisfacție al unei ajustări prin valoarea respectivă a lui χ^2 (§ 87). Să revenim la relația (84). Dacă asociația între cele două variabile este nulă, atunci:

$$(AB) = \frac{(A)\cdot(B)}{N}$$

şi, prin urmare, tipul tabelului care nu ne poate indica nici o asociație între A şi B va fi de forma:

Al doilea atribut	Primul atribut		
	A	α	Total
B	$\frac{(A)(B)}{N}$	$\frac{(B)(\alpha)}{N}$	(B)
β	$\frac{(A)(\beta)}{N}$	$\frac{(\alpha)(\beta)}{N}$	(β)
Total	(A)	(α)	N

Cu cât frecvențele (AB), $(A\beta)$, $(B\alpha)$, $(\alpha\beta)$ se vor depărta mai mult de frecvențele tabelului precedent, cu atât asociația între caracterele A şi B va fi mai accentuată. Vom avea, aşadar, de stabilit în definitiv discordanța între două grupe a câte 4 frecvențe. După formula (53), vom avea:

$$\chi^2 = \frac{\delta_{AB}^2}{\frac{(A)(B)}{N}} + \frac{\delta_{A\beta}^2}{\frac{(A)(\beta)}{N}} + \frac{\delta_{B\alpha}^2}{\frac{(B)(\alpha)}{N}} + \frac{\delta_{\alpha\beta}^2}{\frac{(\alpha)(\beta)}{N}} \qquad (87)$$

unde:

$$\delta_{AB} = -\delta_{A\beta} = -\delta_{B\alpha} = \delta_{\alpha\beta} = \frac{1}{N}\left[(AB)(\alpha\beta) - (A\beta)(B\alpha)\right].$$

Diferențele poartă numele de *contingențele* cazelor respective şi sunt aceleaşi cu cele considerate în cazul coeficientului de asociație.

După un calcul uşor de urmărit, expresia (87) se pune sub forma mai simplă:

$$\chi^2 = \frac{\left[N(AB)(\alpha\beta) - (AB)(B\alpha)\right]^2}{(A)(B)(\alpha)(\beta)} \qquad (88)$$

care a fost numit de *K. PEARSON, pătratul de contingență*. Cu cât χ^2 va fi mai mare cu atât asociația între caracterele A şi B va fi mai pronunțată. Iată deci o măsură a intensității asociației dintre două caractere.

Scanned with OKEN Scanner

Vom trece acum de la valoarea lui χ^2 la valoarea lui P întocmai ca și în cazul coeficientului de satisfacție al unei ajustări. Valoarea lui P în acest caz va corespunde lui $p' = 2$,[1] deci (54):

$$P = \frac{\int_{\chi}^{\infty} e^{-\frac{\chi^2}{2}} d\chi}{\int_{0}^{\infty} e^{-\frac{\chi^2}{2}} d\chi} = \sqrt{\frac{2}{\pi}} \int_{\chi}^{\infty} e^{-\frac{\chi^2}{2}} d\chi \qquad (89)$$

valoare care se poate calcula fie cu ajutorul Tablelor II de la sfârșitul cărții, $P = 2[1 - S_1(\chi)]$, fie cu cele din *Tables for Statisticians and Biometricians*, Part I, p. 2.

TABELUL 102

Suprafața fructului	Culoarea florii		
	Violetă	Albă	Total
Aspră	$\frac{68 \cdot 59}{83} = 48,3$	$\frac{68 \cdot 24}{83} = 19,7$	68
Netedă	$\frac{15 \cdot 59}{83} = 10,7$	$\frac{15 \cdot 24}{83} = 4,3$	15
Total	59	24	83

Să ne referim la tabelul 96. Vom avea, fie aplicând formula (88), fie calculând pe χ^2 prin comparația directă cu tabelul 102 corespunzând clasificării atributelor independente:

$$\chi^2 = 0,708045 \qquad \chi = 0,841$$

---

[1] *Asupra valorii lui p', pentru acest caz, s-a produs o divergență între statisticienii englezi.* Prof. Pearson *susține că valoarea corespunzătoare a lui P trebuie calculată pentru p' = 4, deoarece la bază* χ^2 *este calculat pornind de la diferențele a patru frecvențe.* R. A. Fisher *a introdus aici noțiunea de* grad de libertate. *Dacă noi comparăm frecvențele* $n'_1, n'_2, \ldots, n'_{p+1}$, *cu* $n_1, n_2, \ldots n_{p+1}$, *atunci* $p' = p+1$. *De fapt, susține Fisher, comparația se face numai pentru p din aceste frecvențe, deoarece cum* $\Sigma n = \Sigma n' = N$, *cea de a (p + 1)-a este complet determinată dacă se fixează celelalte. Fisher numește numărul frecvențelor care pot fi fixate arbitrar* gradul de libertate al clasificării; *în cazul citat, gradul de libertate este p.*

Nu este acum greu să ne dăm seama că, în cazul unei 2×2 *clasificări, dacă totalurile marginale sunt fixate, gradul de libertate al distribuției este 1, căci numai una din frecvențele interioare poate fi fixată arbitrar. Deci p = 1 și, prin urmare, p' = 2. Pearson nu găsește justificat însă acest punct de vedere și recomandă p' = 4. Cu toată autoritatea profesorului de la University College din Londra, precum și cu toată deosebita considerație pe care personal o am pentru D-sa, nu mă pot împiedica de a nu fi de partea lui R. A. Fisher.*

Scanned with OKEN Scanner

și deci:

$$P = 0{,}40034$$

ceea ce înseamnă că distribuția din tabelul 96 nu poate fi considerată ca diferind din punct de vedere al teoriei eșantioanelor de cea din tabelul 102, adică de o distribuție în care caracterele în raport cu care s-a făcut clasificarea sunt absolut independente.

Să aplicăm același procedeu clasificării din tabelul 103; vom avea:

$$\chi^2 = 175{,}76 \qquad \chi = 13{,}3.$$

Această valoare este prea mare pentru ca să mai putem calcula pe P cu ajutorul tablelor menționate. De altfel, în acest caz, P are o valoare prea mică pentru ca să mai poată fi obținută pe aceeași cale. Vom recurge atunci la valoarea *log P* care se poate deduce din tablele corespunzătoare (*Tables for Statisticians and Biometricians,* Part I, p.11[1]).

Obținem în cazul nostru:

$$\log P = -39{,}68087 = \overline{40}{,}31913$$

și deci:

$$P = \frac{2085}{10^{40}}\ ^{2}.$$

TABELUL 103

Asociația dintre vaccinarea contra variolei și vindecare

	Vindecare	Deces	Total
Cicatrice prezentă	1.562	42	1.604
Cicatrice absentă	383	94	477
Total	1.945	136	2.081

Reprodus din K. Pearson, ***On a Novel Method of Regarding Association,*** *Drapers' Company Research Memoirs, Biometric Series VIII.*

Probabilitatea ca atributele din tabelul 103 să fie independente este așa de mică încât putem afirma cu certitudine că există o legătură între vaccinarea contra variolei și vindecarea în caz de contractare a bolii.

---

[1] *Prof. K. Pearson dă în aceeași publicație o tablă mai detaliată decât aceasta pentru valorile lui log P, pornind însă, în conformitate cu punctul său de vedere, de la p' = 4.*

[2] *Adică 2085 precedat de 36 de zero până la virgula zecimală.*

Scanned with OKEN Scanner

Gradul acestei legături este măsurat de valoarea lui *P*, care va fi cu atât mai mic cu cât asociaţia va fi mai accentuată. Din nenorocire, după cum spune însuşi prof. K. PEARSON, noi nu putem gândi în milioane şi, prin urmare, valoarea lui *P* nu poate fi un mijloc comod de a aprecia gradul de asociaţie. Acest fapt a făcut pe PEARSON să caute alte procedee de soluţionare a acestei probleme aşa de importante pentru aplicaţiile statisticii matematice[1].

132. *Coeficientul de contingenţă*. Un alt mijloc care să ne permită evaluarea gradului de legătură între două variabile a fost imaginat tot de prof. K. PEARSON, pornind mereu de la semnificaţia diferenţei (84) [2]. Să considerăm distribuţia unei colectivităţi în raport cu diferitele aspecte a două atribute. Vom obţine o distribuţie care se va prezenta sub o formă asemănătoare tabelului 101.

SCHEMA UNUI TABEL DE CONTINGENŢĂ

	A_i		
B_j	(A_iB_j)		(B_j)
	(A_i)		(N)

O astfel de prezentare a unei distribuţii statistice în raport cu două atribute poartă numele de *tabel de contingenţă*. Acesta cuprinde un număr oarecare de coloane verticale şi de linii orizontale care, prin intersecţia lor, formează câte o cază. Fiecare cază este formată din elementele distribuţiei care prezintă în acelaşi timp caracterele A_i şi B_j. Vom nota frecvenţa acestor elemente prin (A_iB_j); totalurile coloanei şi liniei corespunzătoare acestei caze vor reprezenta frecvenţele cazurilor ce prezintă caracterul A_i indiferent dacă-l posedă şi pe B_j sau nu şi viceversa. Vom nota aceste totaluri prin (A_i) şi (B_j).

Contingenţa unei caze va fi reprezentată de diferenţa:

$$\delta_{ij} = (A_iB_j) - \frac{(A_i)(B_j)}{N}. \qquad (90)$$

[1] *A se vedea mai departe Cap. XII.*

[2] *K. Pearson,* ***On the Theory of Contingency and its Relation to Association and Normal Correlation****, Drapers' Company Research Memoirs, Biometric Series I.*

Scanned with OKEN Scanner

Observăm că (A_iB_j) este frecvența statistică a cazei considerate, iar $\frac{(A_iB_j)}{N}$ frecvența teoretică în cazul când atributele respective ar fi complet independente. În cazul independenței perfecte dintre atribute, toate diferențele (90) vor fi nule.

Pornind de la diferențele δ_{ij} , vom calcula ca mai sus:

$$\chi^2 = \sum \frac{\delta_{ij}^2}{\frac{(A_i)(B_j)}{N}} = N\left[\sum \frac{(A_iB_j)^2}{(A_i)(B_j)} - 1\right] \tag{91}$$

valoare egală în cazul unei 2 × 2 clasificări cu (88).

Pornind de aici, PEARSON definește *pătratul mediu de contingență:*

$$\varphi^2 = \frac{\chi^2}{N} = \sum \frac{(A_iB_j)^2}{(A_i)(B_j)} - 1 \tag{92}$$

și apoi *coeficientul de contingență:*

$$C = \sqrt{\frac{\varphi^2}{1+\varphi^2}}. \tag{93}$$

C variază evident de la 0 până la 1, permițându-ne să apreciem pe o scară fixă și ușor de cuprins cu mintea, gradul de asociație a două variabile. Într-adevăr, $C = 0$ atunci și numai atunci când toate diferențele δ_{ij} sunt nule, adică în cazul independenței absolute. În ceea ce privește limita superioară a lui C, trebuie să remarcăm că, deși aceasta este teoretic egală cu 1, această valoare nu este atinsă în practică chiar dacă avem de a face cu o contingență perfectă: orice A_i este B_i și reciproc. Într-adevăr, considerând tabelul 104 corespunzător acestui caz:

TABELUL 104

Contingență perfectă

	A_1	A_2	A_3		A_p	
B_1	n_1	0	0		0	n_1
B_2	0	n_2	0		0	n_2
B_3	0	0	n_3		0	n_3
B_p	0	0	0		n_p	n_p
	n_1	n_2	n_3		n_p	N

Scanned with OKEN Scanner

vom avea:

$$\chi^2 = \sum \frac{n_i^2}{n_i^2} - 1 = p - 1$$

și deci:

$$C = \sqrt{\frac{p-1}{p}} < 1.$$

Aceasta înseamnă că în cazul unui tabel cu 2 linii și 2 coloane cea mai mare valoare ce se poate obține pentru $C = \sqrt{\frac{1}{2}} = 0{,}707$, pe când în cazul unei 3×3 clasificări această valoare este $\sqrt{\frac{2}{3}} = 0{,}816$ etc.

Diferența între aceste valori maxime ale coeficientului de contingență de la caz la caz a dat naștere la unele obiecții împotriva acestui procedeu de măsură a gradului de asociație. PEARSON răspunde însă, și cu drept cuvânt, că acest fapt constituie, din contră, un avantaj al metodei. Într-adevăr, o 2×2 clasificare nu ne poate indica singură o asociație perfectă; s-ar putea întâmpla ca o mai amănunțită clasificare să ne arate că asociația nu este absolut perfectă și, prin urmare, valoarea maximă a lui C, *0,707*, corespunde nivelului până la care se poate stabili gradul de asociație cu ajutorul unei 2×2 clasificări.

TABELUL 105

Asociația între profesia mamei înainte și după naștere

Profesia după naștere	Profesia mamei înainte de naștere		
	Cu profesie	Fără profesie	Total
Cu profesie	750	-	750
Fără profesie	-	250	250
Total	750	250	1000

Vom ilustra acest lucru pe exemplul dat de însuși profesorul PEARSON[1]. Este vorba despre asociația între profesia mamei înainte și după nașterea unui copil. Dacă se disting numai categoriile «cu profesie» și «fără profesie», atunci obținem clasificarea din tabelul 105, pentru care $C = 0{,}707$.

[1] *K. PEARSON, **On a Novel Method of Regarding Association,** Drapers' Company Research Memoirs, Biometric Series VIII, p. 10.*

Scanned with OKEN Scanner

De fapt, tabelul prezintă o asociație perfectă, ceea ce ar însemna că nașterea unui copil nu conduce la o schimbare în profesia mamei. Cu toate acestea, procedându-se la o clasificare mai amănunțită, tabelul 106, observăm că asociația nu mai este perfectă și, prin urmare, nu mai putem afirma că nașterea unui copil nu determină o schimbare în profesia mamei. Faptul că, pentru tabelul 105, $C < 1$ trebuie privit ca unul din meritele coeficientului de contingență, mai ales dacă ținem seama că pentru clasificarea din tabelul 106 $C = 0{,}755$, valoare care nu se depărtează prea mult de prima.

TABELUL 106

Profesia mamei după naștere	Profesia mamei înainte de naștere				
	Lucrătoare în fabrică	Servitoare cu ziua	Lucrătoare acasă	Fără profesie	Total
Lucrătoare în fabrică	540	-	-	-	540
Servitoare cu ziua	58	20	10	-	88
Lucrătoare acasă	70	30	22	-	122
Fără profesie	-	-	-	250	250
Total	668	50	32	250	1000

Tabelul 105

Calculul lui C, adică a lui φ^2, se poate face prin simpla aplicare a formulei (92) sau direct prin comparația clasificării statistice cu aceea care ar corespunde aceleiași probleme, însă în care independența ar fi absolută. Frecvențele corespunzătoare acesteia din urmă se calculează cu ajutorul formulei despre care am vorbit: $\frac{(A_i)(B_j)}{N}$. Calculul lui φ^2 este însă mai anevoios pe această cale și nu este, prin urmare, de recomandat.

Să considerăm tabelul 101 și să aplicăm primul procedeu de calcul al lui φ^2. Vom avea[1]:

$$\begin{aligned} 30{,}5^2 &/ (81{,}5 \times 81{,}5) = 0{,}1401 \\ 2\times\ 23^2 &/ (81{,}5 \times 665) = 0{,}0195 \\ 2\times\ 16^2 &/ (81{,}5 \times 674{,}5) = 0{,}0093 \\ 2\times\ 12^2 &/ (81{,}5 \times 525{,}5) = 0{,}0067 \\ 416^2 &/ (665 \times 665) = 0{,}3913 \\ 2\times 158^2 &/ (665 \times 674{,}5) = 0{,}1113 \end{aligned}$$

[1] *Din cauza simetriei tabelului considerat, am grupat la un loc termenii pe care îi întâlnim de două ori în suma* $\sum \frac{(A_iB_j)}{(A_i)(B_j)}$.

$$
\begin{aligned}
2\times\ 67{,}75^2 &/(665 \times 525{,}5) = 0{,}0263\\
2\times\ 0{,}25^2 &/(665 \times 37{,}5) = 0{,}0000\\
394^2 &/(674{,}5 \times 674{,}5) = 0{,}3412\\
2\times\ 98{,}25^2 &/(674{,}5 \times 525{,}5) = 0{,}0545\\
2\times\ 8{,}25^2 &/(674{,}5 \times 37{,}5) = 0{,}0054\\
328{,}5^2 &/(525{,}5 \times 525{,}5) = 0{,}3908\\
2\times\ 19^2 &/(525{,}5 \times 37{,}5) = 0{,}0366\\
10^2 &/(37{,}5 \times 37{,}5) = 0{,}0711
\end{aligned}
$$

$$\sum \frac{(A_i B_j)}{(A_i)(B_j)} = 1{,}6041$$

$$\varphi^2 = 0{,}6041$$

$$C = \sqrt{\frac{0{,}6041}{1{,}6041}} = \sqrt{0{,}3766} = 0{,}61.$$

Dacă atributele considerate ar fi fost absolut independente, atunci clasificarea pentru același număr de cazuri s-ar fi prezentat sub forma din tabelul 107.

TABELUL 107

Clasificarea în cazul independenței ideale a atributelor din tabelul 101

Al 2-lea frate	Primul frate: Roșu	Blond	Castaniu	Închis	Negru	Total
Roșu	3,4	27,3	27,7	21,6	1,5	81,5
Blond	27,3	212,8	226,1	176,2	12,6	665
Castaniu	27,7	226,1	229,3	178,7	12,7	674,5
Închis	21,6	176,2	178,7	139,1	9,9	525,5
Negru	1,5	12,6	12,7	9,9	0,8	37,5
Total	81,5	665,0	674,5	525,5	37,5	1 984,0

Coeficientul de contingență, după opiniile celor mai de seamă statisticieni, sprijinite de altfel pe însăși teoria matematică a lui C, nu trebuie aplicat pentru măsura legăturii dintre două atribute decât dacă tabelul de contingență are cel puțin 5 linii și 5 coloane. La nevoie, el poate fi aplicat și clasificărilor mai grosolane, însă cu toată rezerva care decurge din încălcarea principiului de mai sus.

Scanned with OKEN Scanner

Bineînţeles că orice calcul de coeficient de contingenţă poate fi completat prin determinarea probabilităţii P ca distribuţia considerată să fie un eşantion al unei colectivităţi în care atributele studiate ar fi independente (§ 131). Valoarea lui P se obţine din tablele cunoscute, adoptând $p' = (l-1)(c-1) + 1$, unde l şi c reprezintă numărul liniilor şi coloanelor din tabelul de contingenţă respectiv. În general însă, valorile lui χ^2 ies afară din tablele calculate până în prezent, aşa că valoarea lui P trebuie calculată direct. Procedeul de calcul este însă foarte anevoios şi munca depusă nu corespunde utilizării valorii aflate[1].

[1] *Procedeul de calcul direct al lui P este arătat în* ***Tables for Statisticians and Biometricians****, Part I, p. XXXIII. Tot în această lucrare găsim câteva tabele care să ne înlesnească această operaţie.*

Scanned with OKEN Scanner

Capitolul XI

Regresie, corelație

133. *Legi rigide și legi stocastice.* În cazul când atributele în raport cu care se face o dublă clasificare sunt măsurabile, studiul legăturii stocastice dintre atribute se poate face dintr-un punct de vedere ce reamintește legea rigidă. Acest procedeu ne va permite să punem în evidență în același timp asemănările dintre legile stocastice și legile rigide, precum și deosebirile dintre aceste două forme sub care apar legile naturale.

Să examinăm, în primul rând și în mod sumar, forma legilor rigide și, pentru a ne fixa ideile, să considerăm legea căderii libere a corpurilor în vid. Avem de-a face în acest caz cu o legătură între două variabile: timpul și spațiul parcurs, ambele considerate din momentul căderii. Presupunând că forma legăturii între acestea două ne este necunoscută, unul din mijloacele cele mai simple de a o descoperi este efectuarea unui șir de experiențe. Rezultatele acestor experiențe pot fi reprezentate grafic pe un sistem de axe de coordonate, purtând timpul pe axa absciselor, iar spațiul pe ordonate. Vom obține astfel în plan un număr de puncte, care, unite printr-o curbă continuă, vor individualiza prin forma acesteia din urmă natura legăturii dintre timp și spațiul parcurs în cazul căderii libere în vid a corpurilor. În legătură cu această lege, trebuie să observăm două lucruri. În primul rând, trebuie relevat faptul că, experiența repetându-se, vom obține aceeași valoare a spațiului pentru același timp, iar în al doilea, că legătura între aceste două variabile se poate exprima algebric prin relația:

$$s = \frac{1}{2} g t^2$$

prin urmare, curba reprezentativă a legii este o parabolă de ordinul 2.

Prima caracteristică a legii considerate, adică faptul că la o valoare a lui t corespunde o singură valoare a lui s în toate experiențele, constituie însăși definiția legilor rigide. Exprimarea într-o formă analitică a legăturii dintre variabile, deși nu este o condiție *sine qua non* a acestor legi, este totuși îndeplinită de la sine pentru aproape totalitatea legilor rigide. Și pentru a ne rezuma, vom spune că avem de a face cu o astfel de lege când unei valori a uneia din variabile îi va corespunde o valoare *unică și bine determinată* pentru

Scanned with OKEN Scanner

cea de-a doua. *Aşadar, legea rigidă este, de fapt, exprimarea unei corespondenţe de la valoare la valoare.*

TABELUL 108

Contingenţa dintre înălţimea tatălui şi înălţimea fiului

Înălţimea fiului = y	Înălţimea tatălui = x [1]									
	58,5-	60,5-	62,5-	64,5-	66,5-	68,5-	70,5-	72,5-	74,5	Total
59,5-61,5	-	-	1,5	1	1	-	-	-	-	3,5
61,5-63,5	0,5	2,75	5,75	9,5	5	0,25	0,25	-	-	24
63,5-65,5	4	7,75	20	41,5	17,25	8,25	1,25	-	-	100
65,5-67,5	2	10	32	73	78,75	33,5	7,25	1	-	237,5
67,5-69,5	-	4,5	27,75	65,5	95	93,25	31,5	4,5	1	323
69,5-71,5	-	-	6,75	38,25	61	77,5	39,5	11	2	236
71,5-73,5	-	-	0,25	5,75	24,75	34,5	32,25	7	0,5	105
73,5-75,5	-	-	1	3	6,25	6,75	13	5,5	2	37,5
75,5-77,5	-	-	-	-	2,5	1,5	1,5	2,5	-	8
77,5-79,5	-	-	-	-	-	2	0,5	1	-	3,5
Total	6,5	25	95	237,5	291,5	257,5	127	32,5	5,5	1078

*Reprodus din K. Pearson, **On the Theory of Contingency**..., p.29.*

Să trecem acum la cazul unei legi stocastice. Urmărind valorile uneia din variabile în raport cu valorile celeilalte, observăm că nu mai avem de-a face cu aspectul rigid de mai sus şi că, pentru aceeaşi valoare a primei variabile, putem obţine valori diferite pentru cea de-a doua. Acest fapt se desprinde şi mai bine din consideraţia tabelului 108.

Constatăm astfel că unei valori[2] a primei variabile îi corespunde o serie întreagă de valori ale celei de-a doua variabile distribuite într-un anumit mod. Aşadar, corespondenţa între variabile nu se mai face ca în cazul legilor rigide de la valoare la valoare, *ci de la valoare la distribuţie.* Ceea ce va varia în raport cu una din variabile va fi, prin urmare, structura distribuţiei corespunzătoare celeilalte şi această variaţie va constitui însăşi legea stocastică. Această variaţie - esenţială pentru o lege stocastică, pentru că dacă n-ar exista ar însemna că nu poate fi vorba despre vreo legătură între cele două variabile - se poate constata în mod vizual pe tabelul 108. Un studiu precis al acesteia şi, prin urmare, al legii stocastice respective va necesita urmărirea variaţiei tuturor valorilor caracteristice ale distribuţiilor respective de-a lungul scărilor de variaţie ale celor două variabile.

---

[1] *Înălţimile sunt măsurate în degete. Când una din înălţimi a fost egală cu limita dintre două intervale, s-a repartizat frecvenţa 0,5 intervalelor alăturate, iar când această egalitate a avut loc pentru ambele înălţimi, s-a repartizat frecvenţa 0,25 celor patru caze adiacente.*

[2] *De fapt, în tabelul considerat nu avem valori individuale, ci o scară de intervale. Acest lucru nu ne împiedică totuşi să facem constatarea de mai sus.*

Scanned with OKEN Scanner

134. *Reprezentarea grafică a legăturilor stocastice.* În ceea ce privește reprezentarea grafică a unei legi stocastice, să considerăm datele din tabelul 109, care, spre deosebire de cele din tabelul 108, ne permit urmărirea legăturii dintre înălțime și greutate prin valori individuale. Graficul 48 este obținut prin aplicarea principiului coordonatelor carteziene, întocmai ca și pentru legile rigide. Valorile înălțimilor sunt purtate pe axa absciselor, iar cele ale greutăților pe ordonate. Fiecărui caz îi corespunde un punct în planul axelor. Vom obține astfel 76 de puncte pe care nu ne mai putem gândi să le unim printr-o linie continuă, deoarece sunt răspândite în plan ca un roi. Aspectul general al acestui ansamblu de puncte ne permite un prim contact rudimentar cu natura legii stocastice. Se observă astfel că răspândirea punctelor nu este uniformă și că există o tendință de concentrare de-a lungul unei linii drepte înclinate spre dreapta[1].

TABELUL 109

Vârsta, înălțimea și greutatea unui grup de copii între 13 și 15 ani[2]

Nr.	Vârsta	Înălțimea	Greutatea	Nr.	Vârsta	Înălțimea	Greutatea
(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	169	59,5	84,75	39	175	59,5	87,5
2	177	58	77	40	157	59,5	77
3	175	65,25	103	41	171	59	88,25
4	167	60	87,75	42	159	52,5	67,5
5	179	63,75	112,5	43	171	63	91,5
6	175	58	70	44	162	61,5	97,25
7	171	58	91	45	170	56,75	84,75
8	177	65	117,25	46	161	62	102
9	179	57,5	82,25	47	164	62,5	95
10	171	59,75	95,75	48	174	64	102,5
11	160	55	71,5	49	176	61,5	94
12	175	63,75	106,75	50	159	56,25	70
13	164	55,75	73	51	156	61,25	91
14	173	62,5	106,75	52	166	59,5	101,25
15	172	60	84	53	176	59,5	93,75

[1] *Ar fi nimerit ca pentru reprezentarea grafică a tabelelor de contingență să adoptăm pentru axele de coordonate același sens ca cel întrebuințat până în prezent și care în matematici poartă numele de* sens direct *(în așezarea obișnuită, sensul pozitiv al absciselor fiind de la stânga spre dreapta, cel al ordonatelor va fi de jos în sus). Cum însă există tradiția de a prezenta aceste tabele cu valorile lui y crescând de sus în jos, este mai logic, pentru a ușura confruntarea directă dintre grafic și tabel, să adoptăm pentru axele primului* sensul retrograd.

[2] *Vârsta este socotită în luni, înălțimea în degete (inch), iar greutatea în pfunzi (pound).*

Scanned with OKEN Scanner

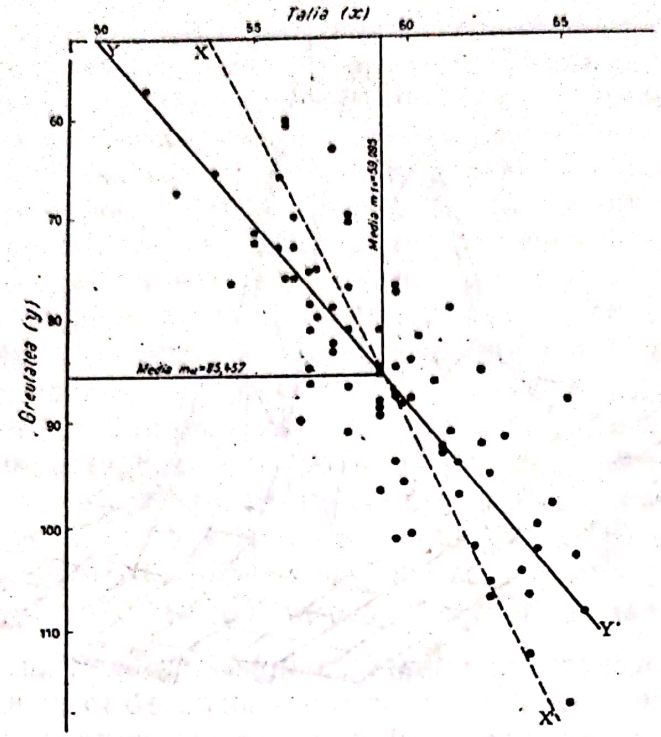
(continuare tab. 109)

(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
16	162	57	76	54	164	53,75	65,5
17	168	62,5	105,5	55	176	64,5	98
18	156	56,25	73	56	168	58	86,5
19	175	59	96,75	57	166	57	75
20	178	61	93	58	158	60,5	88
21	165	59	81	59	162	55	72,5
22	178	56,75	81	60	156	57	79,75
23	160	57,5	78,75	61	159	56,75	75,25
24	157	62,25	85	62	161	56,5	89,75
25	168	54,25	76,5	63	166	65,5	108,5
26	174	63,5	104,5	64	157	56,75	86,25
27	171	60,25	81,75	65	166	64	100
28	165	57	81	66	168	57	70,5
29	160	61,25	79	67	159	55	70
30	172	61	92,5	68	164	59	85,25
31	178	57,5	63,25	69	163	59,75	88,25
32	158	57,5	83,25	70	167	56	60,5
33	169	56,75	78,5	71	167	56	60,5
34	166	62,25	92,25	72	158	55,75	66
35	168	59,5	77	73	168	59	89
36	172	59	89	74	159	51,5	57,25
37	176	59	94,75	75	159	60	100,75
38	163	60,75	86	76	163	56,25	76

*Extras din A. Bowley, **Eléments de statistique**, Paris, Giard, 1929, p. 135.*

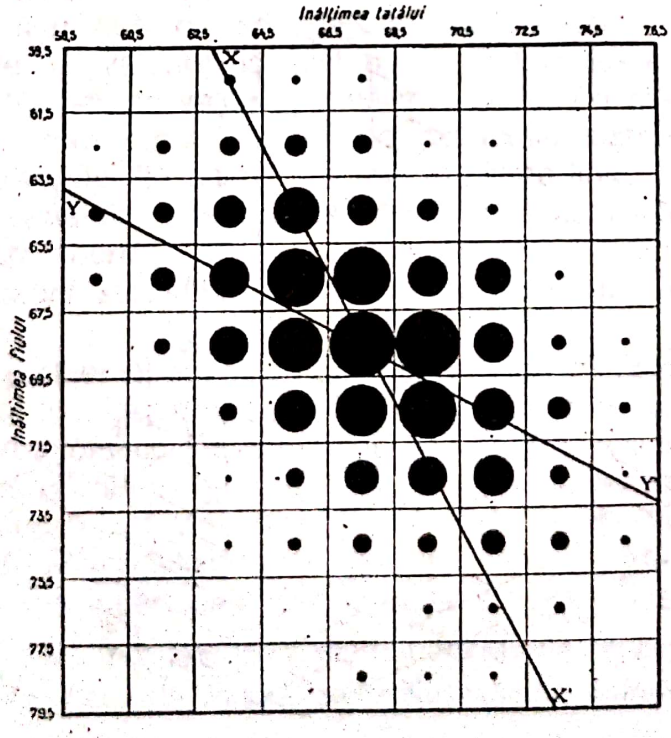
Dacă legătura între două variabile este evidenţiată printr-un tabel de contingenţă în care clasele sunt formate din intervale, cum este cazul tabelului 108, reprezentarea grafică a distribuţiei se poate face în două moduri. Primul constă din reprezentarea frecvenţelor tabelului de contingenţă prin cercuri de suprafaţă proporţională cu acestea. Poziţia acestor cercuri va fi fixată în raport cu două axe de coordonate corespunzând celor două variabile, coordonatele centrului unui astfel de cerc vor fi determinate de mijlocul intervalelor corespunzătoare cazei respective din tabelul de contingenţă. Referindu-ne la tabelul 108 şi la frecvenţa cazei pentru care înălţimea tatălui este cuprinsă între 70,5-72,5, iar a fiului între 65,5-67,5, aceasta va fi reprezentată printr-un cerc de suprafaţă proporţională cu 7,25 şi cu centrul în punctul de coordonate 71,5 şi 66,5. Se obţine prin acest procedeu graficul 49. Se observă şi pe acest grafic cum cercurile reprezentând frecvenţele se concentrează ca număr şi suprafaţă de-a lungul unei linii.

Scanned with OKEN Scanner



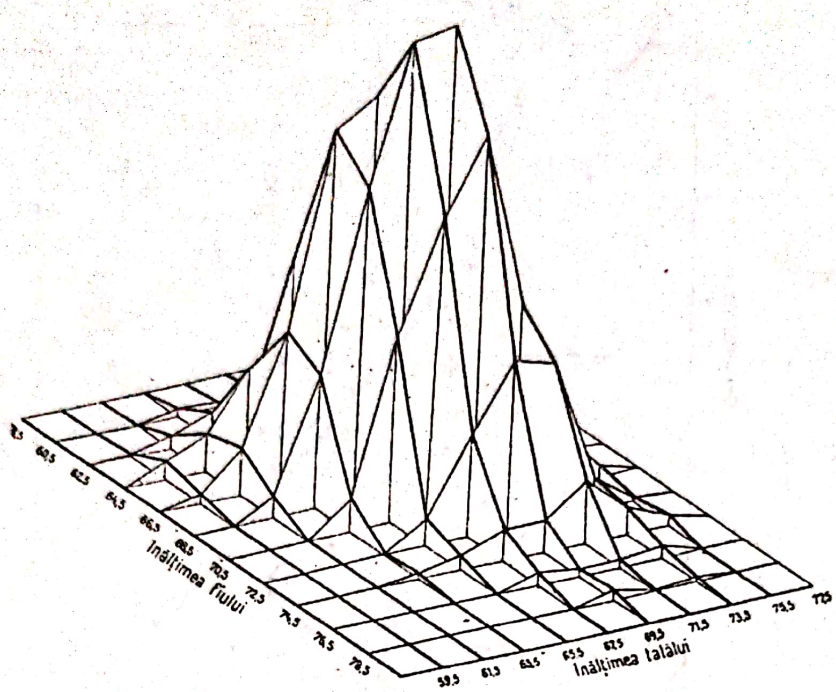
Tabelul 109

Graficul 48. Reprezentarea grafică a unei serii de valori legate stocastic.



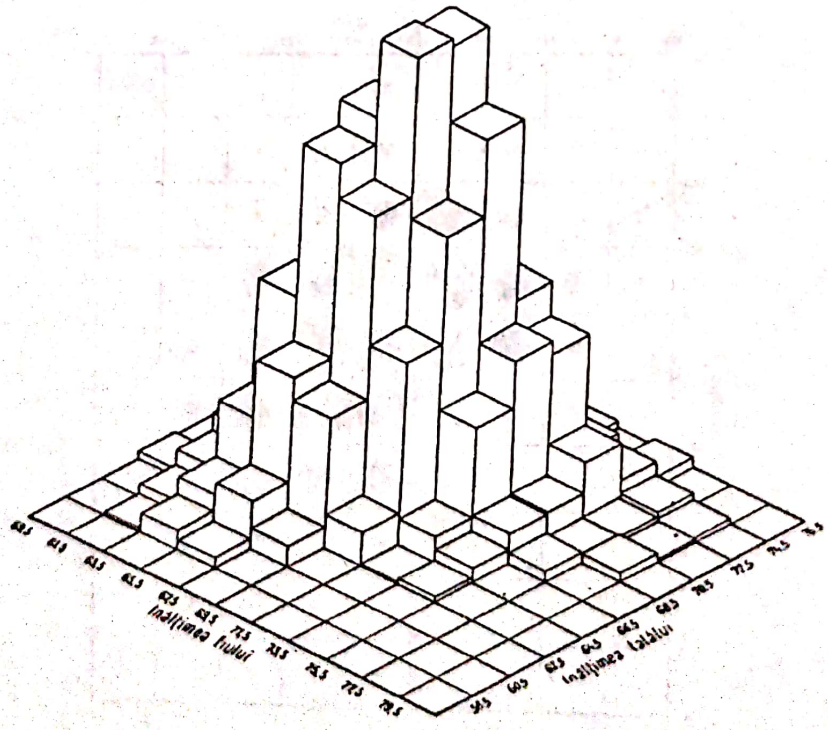
Tabelul 108

Graficul 49. Reprezentarea grafică a unui tabel de contingență.



Tabelul 108

Graficul 50. Reprezentarea grafică a unei distribuții de două variabile corelate.



Tabelul 108

Graficul 51. Reprezentarea grafică a corelației dintre două variabile.

Scanned with OKEN Scanner

Al doilea mod de a reprezenta grafic un tabel de contingenţă constă din reprezentarea frecvenţelor prin segmente proporţionale cu acestea şi ridicate perpendicular pe planul axelor de coordonate în centrul fiecărei caze.

Prin unirea extremităţilor segmentelor consecutive se obţine o reţea de linii poligonale care ne dau impresia unei suprafeţe poliedrale (graficul 50). Acest mod de reprezentare grafică a unei distribuţii în raport cu două variabile corespunde poligonului de frecvenţă al unei distribuţii în raport cu o singură variabilă. Tot ceea ce s-a spus mai sus în legătură cu poligonul de frecvenţă rămâne valabil, mutatis mutandis, şi pentru *poliedrul de frecvenţe.*

Frecvenţele unui tabel de contingenţă mai pot fi reprezentate şi prin paralelipipede de volum proporţional cu acestea şi având ca bază pătratul sau dreptunghiul corespunzător cazei respective (graficul 51).

Această reprezentare grafică este în totul analoagă histogramei. Bineînţeles că ambele reprezentări ne conduc în mod firesc la concepţia unei *suprafeţe de corelaţie*, limita teoretică a poliedrului de frecvenţe având o formă regulată, bucurându-se de aceleaşi proprietăţi ca şi curba de frecvenţe şi prezentând acelaşi interes de cercetare ca şi aceasta din urmă[1].

135. *Aspectul general al studiului legăturilor stocastice.* Prin reprezentarea grafică a unui tabel de contingenţă nu se pune în evidenţă decât simplul aspect global al legăturii stocastice dintre variabilele corespunzătoare. Un studiu adevărat ştiinţific al unei astfel de legături nu se poate limita numai la reprezentarea grafică, procedeu subiectiv şi care poate conduce la concluzii diferite de la un cercetător la altul. Este absolut necesar deci să construim o metodă analitică numerică adaptată acestui scop.

Am văzut mai sus că o lege stocastică trebuie concepută ca o corespondenţă de la valoare la distribuţie. Referindu-ne la tabelul 108, observăm că pentru fiecare valoare a înălţimii tatălui (x) avem o distribuţie a înălţimii fiului (y). Vom numi o astfel de distribuţie *distribuţie legată,* spre deosebire de distribuţiile din ultima coloană şi ultima linie, care cuprind toate elementele unei variabile independent de valorile celeilalte şi care poartă numele de *distribuţii libere.*

Studiul legăturii stocastice va fi, prin urmare, un studiu al variaţiei distribuţiilor legate.

Astfel pusă problema, aceasta nu poate fi rezolvată complet decât în mod teoretic. Rezultatele la care se ajunge pe această din urmă cale nici nu prezintă vreun interes de ordin practic. De aceea, noi ne vom mulţumi să reducem studiul legilor stocastice la o analiză a variaţiei valorilor caracteristice ale distribuţiilor legate şi încă din aceste valori nu vom considera decât pe cele mai importante.

Vom ajunge pe această cale la descompunerea legii stocastice într-un anumit număr de corespondenţe de la valoare la valoare. Într-adevăr, să însemnăm cu $m_{01}^{(i)}$ media aritmetică a distribuţiei legate corespunzătoare valorii x_i a primei variabile. Această medie va fi dedusă din valorile variabilei y din

[1] *Figurile geometrice în spaţiu la care dă naştere o astfel de reprezentare a unui tabel de contingenţă poartă numele de* stereograme.

distribuţia legată respectivă şi se va numi *media legată a lui y*, spre deosebire de media distribuţiei libere pe care o însemnăm prin m_{01} şi care poartă numele de *medie liberă*. Fiecărei valori a lui *x* îi va corespunde o medie legată pentru *y*; obţinem astfel o primă corespondenţă între valori. Evident, o corespondenţă analoagă poate fi concepută între diferitele valori y_j şi mediile legate $m_{10}^{(j)}$. Mergând mai departe, putem considera momentele de diferite ordine ale distribuţiilor legate $m_{02}^{(i)}, m_{03}^{(i)}$, ... şi stabili o corespondenţă între valoarea *x* şi valorile fiecăruia din aceste momente. Ansamblul acestor corespondenţe va forma o imagine analitică a legii stocastice respective.

136. *Liniile de regresie.* Să considerăm distribuţia din tabelul 108. Calculând fiecare medie legată a lui *y* (înălţimea fiului) şi înlocuind fiecare interval printr-o valoare unică egală cu mijlocul lui, obţinem corespondenţa din tabelul 110.

TABELUL 110

Înălţimea tatălui (x_i)	59,5	61,5	63,5	65,5	67,5	69,5	71,5	73,5	75,5
Media legată $m_{01}^{(i)}$	64,96	65,80	66,72	67,40	68,54	69,53	70,72	71,92	71,78

Tabelul 108

Media liberă a lui *y* se poate calcula fie direct, pornind de la frecvenţele din ultima coloană a tabelului 108, fie în mod indirect, făcând media mediilor legate $m_{01}^{(i)}$ şi atribuind fiecăreia dintre acestea o frecvenţă egală cu numărul elementelor ce o compun[1]. Notând prin n_{ij} frecvenţa corespunzătoare cazei (x_i, y_j) şi prin $n_{i|}$, $n_{|j}$ frecvenţele aceloraşi valori în distribuţiile libere, vom avea:

$$\sum_i \sum_j n_{ij} = N \qquad \sum_i n_{ij} = n_{|j} \qquad \sum_j n_{ij} = n_{i|}$$

Proprietatea de care vorbim se traduce algebric prin relaţiile:

$$\frac{\sum n_{i|}\, m_{01}^{(i)}}{N} = m_{01} \qquad \frac{\sum n_{|j}\, m_{10}^{(j)}}{N} = m_{10}$$

Se obţine pe această cale, pentru tabelul 110, $m_{01} = 68{,}66$.

Legătura între mediile legate şi valorile variabilei *x* se poate reprezenta grafic pe un sistem de axe de coordonate. Unind punctele rezultând din reprezentarea diferitelor perechi de valori x_i, $m_{01}^{(i)}$, obţinem o linie poligonală numită după Sir FRANCIS GALTON, *linia de regresie a lui y în x*. Linia de regresie obţinută prin acest procedeu reprezintă numai în limitele fluctuaţiilor datorate hazardului linia de regresie ideală, pe care, din cauza formei de variaţie continuă şi regulată

[1] *Acest din urmă procedeu este bazat pe o proprietate a mediei aritmetice expusă mai sus (§ 46). Calculul mediei libere pe ambele căi poate fi considerat ca un mijloc de verificare a operaţiilor.*

Scanned with OKEN Scanner

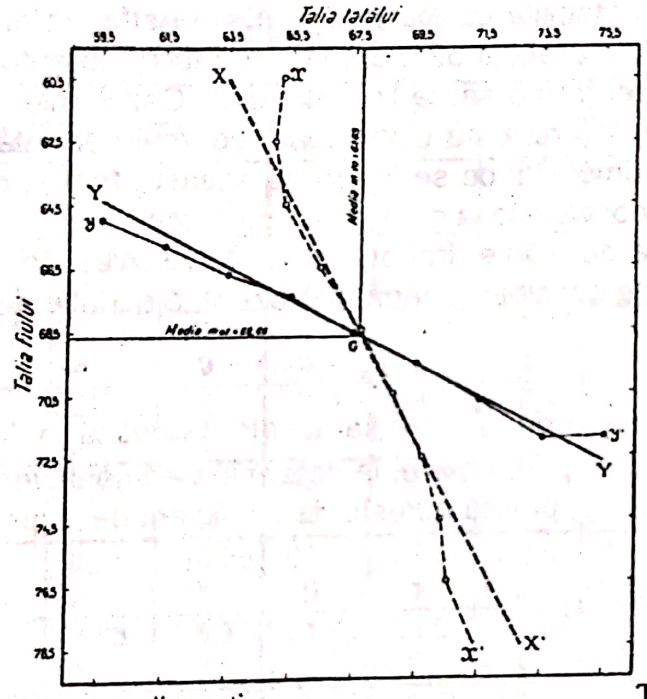
pe care o îmbracă mai toate legile naturale, noi nu o putem concepe altfel decât ca o linie cu alură regulată[1]. Bineînțeles că, după cum există o linie de regresie a lui *y* în *x*, vom avea o linie de regresie și pentru *x* în *y* (tabelul 111, graficul 52). Nimic însă nu ne spune că aceste linii trebuie să se confunde. Dimpotrivă, *existența a două linii de regresie caracterizează complet legile stocastice, confundarea acestora într-una singură caracterizând, pe de altă parte, legile rigide.*

TABELUL 111

Înălțimea fiului (y_j)	60,5	62,5	64,5	66,5	68,5	70,5	72,5	74,5	76,5	78,5
Media legată ($m_{10}^{(j)}$)	65,22	64,96	65,30	66,46	67,74	68,74	69,70	70,28	70,50	70,92

Tabelul 108

$m_{10} = 67,69$



Tabelele 110, 111, 113

XX', YY' = regresii teoretice
xx', yy' = regresii statistice

Graficul 52 . Liniile de regresie ale distribuției înălțimilor părinților și descendenților.

[1] *Se observă din tabelul 110 și din graficul corespunzător că înălțimea medie a fiului pentru o înălțime dată a tatălui este mai mare decât aceasta în cazul valorilor pentru care $x < m_{10}$ (67,69) și mai mică în cazul contrar. Această lege a ereditații, pusă în evidență pentru prima oară de Sir Francis Galton, ne arată că ereditatea caută să corecteze abaterile față de medie; din părinți a căror talie este mai mare decât media grupului se nasc copii cu o talie inferioară lor și reciproc. Există, așadar, o regresie către valoarea medie, de unde și numele liniei reprezentând legătura între una din variabile și mediile legate. Termenul de linie de regresie este însă impropriu pentru cazul general.*

Scanned with OKEN Scanner

În acest din urmă caz, obținem linia reprezentând legătura între cele două variabile. Prin urmare, cunoașterea formei analitice a legăturii lui *y* față de *x*, sub forma de regresie, nu ne permite în cazul legăturii stocastice să determinăm forma legăturii lui *x* față de *y*. Acest lucru, care se poate face pentru legile rigide prin operația cunoscută în matematică sub numele de inversiunea funcțiilor, este o imposibilitate în domeniul fenomenelor colective.

137. *Regresia liniară*. Linia de regresie teoretică a unei distribuții în raport cu două variabile poate avea formele cele mai variate. Această formă ne este însă necunoscută, deoarece, după cum am spus de repetate ori, noi nu avem posibilitatea de a cunoaște colectivitatea ideală. Putem calcula linia de regresie ideală; rezultatele însă nu vor fi valabile decât privite sub unghiul fluctuațiilor datorate hazardului. În afară de acest lucru, este absolut necesar ca să avem o idee prealabilă asupra formei acestei linii, calculul diferind de la o formă la alta. Forma liniei de regresie ideală poate fi dedusă întrucâtva din aceea a liniei de regresie statistică stabilită ca mai sus (§ 136). Astfel, referindu-ne la distribuția din tabelul 108 și la graficul 52, suntem pe deplin îndreptățiți să căutăm o linie de regresie a cărei formă să fie linia dreaptă. Cazul liniei de regresie dreaptă, adică al *regresiei liniare*, este dintre toate cel mai des întâlnit, cel mai ușor de tratat din cauza simplității ce se leagă de această formă, dar în același timp și cel mai bogat în concluziile la care poate conduce.

Fiind vorba de o linie dreaptă, aceasta va avea, în raport cu sistemul de axe de coordonate considerat până în prezent, o ecuație de forma:

$$\tilde{y}_x = b_{21}x + c \tag{94}$$

unde am însemnat prin $\tilde{y}_x$ media legată teoretică a lui *y* în raport cu *x*. Problema se reduce, prin urmare, la determinarea coeficienților b_{21} și *c*.

Coeficientul b_{21} poartă numele de coeficient de regresie a lui *y* în *x* și este dat de formula:

$$b_{21} = \frac{\mu_{11}}{\mu_{20}} \tag{95}$$

unde μ_{20} este fluctuația liberă a lui *x*, iar μ_{11}, momentul-produs în raport cu mediile libere ale variabilelor. Formula acestuia este:

$$\mu_{11} = \frac{\sum n_{ij}(x_i - m_{10})(y_j - m_{01})}{N} \tag{96}$$

sau, dacă însemnăm cu ξ_i, η_j abaterile acestor valori în raport cu mediile respective, vom avea:

$$\mu_{11} = \frac{\sum n_{ij}\xi_i\eta_j}{N} \tag{96 bis}$$

Scanned with OKEN Scanner

Analogia lui μ_{11} cu forma momentelor distribuţiilor heterograde de o singură variabilă este evidentă.

Pentru calcularea acestui moment, vom face uz de metoda cunoscută a schimbării originii şi de formula corespunzătoare acestui caz:

$$\mu_{11} = \left[m_{11}(a,b) - m_{10}(a)m_{01}(b)\right] \times d_1 d_2 \qquad (97)$$

în care *a* şi *b* sunt originile arbitrare, d_1, d_2 - noile unităţi de măsură, iar m_{11} *(a, b)* - momentul calculat în raport cu aceste origini şi cu aceste unităţi de măsură:

$$m_{11} = \frac{\sum n_{ij}\alpha_i\beta_j}{N} \qquad (98)$$

α_i, β_j fiind respectiv abaterile lui x_i, y_j în raport cu *a*, *b*.

TABELUL 112 A

Înălţimea fiului	Înălţimea tatălui									
	59,5	61,5	63,5	65,5	67,5	69,5	71,5	73,5	75,5	Total
60,5	-	-	1,5	1	1	-	-	-	-	3,5
	-	-	+8	+ 4	0	-	-	-	-	- 4
62,5	0,5	2,75	5,75	9,5	5	0,25	0,25	-	-	24
	+12	+9	+6	+ 3	0	- 3	- 6	-	-	- 3
64,5	4	7,75	20	41,5	17,25	8,25	1,25	-	-	100
	+ 8	+6	+4	+ 2	0	- 2	- 4	-	-	- 2
66,5	2	10	32	73	78,75	33,5	7,25	1	-	237,5
	+ 4	+3	+2	+ 1	0	- 1	- 2	- 3	-	- 1
68,5	-	4,5	27,75	65,5	95	93,25	31,5	4,5	1	323
	-	0	0	0	0	0	0	0	0	0
70,5	-	-	6,75	38,25	61	77,5	39,5	11	2	236
	-	-	- 2	- 1	0	+ 1	+ 2	+ 3	+ 4	+ 1
72,5	-	-	0,25	5,75	24,75	34,5	32,25	7	0,5	105
	-	-	- 4	- 2	0	+ 2	+ 4	+ 6	+ 8	+ 2
74,5	-	-	1	3	6,25	6,75	13	5,5	2	37,5
	-	-	- 6	- 3	0	+ 3	+ 6	+ 9	+12	+ 3
76,5	-	-	-	-	2,5	1,5	1,5	2,5	-	8
	-	-	-	-	0	+ 4	+ 8	+12	-	+ 4
78,5	-	-	-	-	-	2	0,5	1	-	3,5
	-	-	-	-	-	+ 5	+10	+15	-	+ 5
Total	6,5	25	95	237,5	291,5	257,5	127	32,5	5,5	1078
	- 4	- 3	- 2	- 1	0	+ 1	+ 2	+ 3	+ 4	-

138. *Calculul coeficienţilor de regresie.* Pentru ilustrarea procedeului de calcul al lui μ_{11}, să ne referim tot la distribuţia din tabelul 108. În acest scop vom forma tabelul 112 A, în care intervalele sunt înlocuite prin valori unice, egale cu

Scanned with OKEN Scanner

mijlocul celor dintâi, iar în fiecare caz este înscrisă, pe lângă frecvența respectivă, și produsul abaterilor $\alpha_i\beta_j$ în raport cu originile alese $a = 67{,}5$, $b = 68{,}5$.

TABELUL 112 B

Calculul momentului μ_{11}

Abateri	Frecvențele corespunzătoare		Produsul (1) × (2)
(1)	(2)		(3)
1	73+77,5-33,5-38,25	= 78,75	78,75
2	32+41,5+34,5+39,5-8,25-7,25-6,75-5,75	= 119,50	239,00
3	10+9,5+6,75+11-0,25-1-3	= 33,00	99,00
4	2+20+1+1,5+32,25+2-1,25-0,25	= 57,25	229,00
5	2	= 2,00	10,00
6	7,75+5,75+13+7-0,25-1	= 32,25	193,50
8	4+1,5+1,5+0,5	= 7,50	60,00
9	2,75+5,5	= 8,25	74,25
10	0,5	= 0,50	5,00
12	0,5+2,5+2	= 5,00	60,00
15	1	= 1,00	15,00
			1063,50

Tabelul 108

$$m_{10}(a) = \frac{102{,}5}{1078} = 0{,}0951 \qquad m_{01}(b) = \frac{84{,}5}{1078} = 0{,}0784$$

$$m_{11}(a,b) = \frac{1063{,}5}{1078} = 0{,}9865$$

$$\mu_{11} = (0{,}9865 - 0{,}0951 \times 0{,}0784) \times 4 = 3{,}9162$$

$$\mu_{20} = \left(\frac{2092{,}5}{1078} - 0{,}0951^2\right) \times 4 = 7{,}7284 \qquad \mu_{02} = \left(\frac{2118{,}5}{1078} - 0{,}0784^2\right) \times 4 = 7{,}8364$$

Corecția Sheppard $\left(\frac{d^2}{12}\right)$ $\frac{0{,}3333}{7{,}3951}$ $\frac{0{,}3333}{7{,}5031}$

Observăm că acest tabel este împărțit de coloana și linia corespunzătoare valorilor alese ca origine în patru cadrane și că, pentru un astfel de cadran, toate produsele $\alpha_i\beta_j$ sunt de același semn. Mai observăm, de asemenea, că acest produs, abstracție făcând de semne, are aceeași valoare pentru mai multe caze. Prin urmare, fixându-ne ideile asupra cazului când $|\alpha_i\beta_j| = 2$, în loc să înmulțim pe rând fiecare din frecvențele corespunzătoare cu factorul 2,

Scanned with OKEN Scanner

vom face suma algebrică a acestor frecvenţe, luându-le cu semnul "+" când abaterea este pozitivă şi cu semnul "-" în caz contrar; apoi vom înmulţi rezultatul cu valoarea absolută a produsului considerat. Acest calcul se poate dispune aşa cum se vede în tabelul 112 B. Suma coloanei (3) împărţită prin suma totală a frecvenţelor ne va da pe m_{11}(a,b). Valorile m_{10}(a), m_{01}(b) se obţin la fel ca în cazul distribuţiilor de o singură variabilă (§ 47).

Momentul μ_{11} se obţine prin aplicarea formulei (97).

Pentru a obţine valoarea lui μ_{20}, vom calcula prin procedeul cunoscut (§ 67) fluctuaţia distribuţiei libere a lui x:

$$\mu_{20} = \left[m_{20}(a) - \overline{m_{10}(a)}^2 - \frac{1}{12} \right] \cdot d^2 = 7{,}3951$$

şi, prin urmare:

$$b_{21} = 0{,}53 .$$

La fel vom avea pentru valoarea coeficientului de regresie a lui x în y:

$$b_{12} = \frac{3{,}9162}{7{,}5031} = 0{,}52.$$

În ceea ce priveşte determinarea celui de-al doilea coeficient din ecuaţia (94), este preferabil să ne servim de faptul că linia de regresie trece în cazul regresiei liniare prin punctul G de coordonate m_{10},m_{01} (centrul de greutate al distribuţiei considerate) şi să o scriem sub forma:

$$\tilde{y}_x - m_{01} = b_{21}(x - m_{10}). \tag{99}$$

La fel vom avea pentru linia de regresie a lui x în y:

$$\tilde{x}_y - m_{10} = b_{12}(y - m_{01}). \tag{99 bis}$$

Cu ajutorul acestor ecuaţii putem determina mediile legate teoretice ale tabelului 108, aşa cum se vede din tabelul 113.

O dată valorile lui $\tilde{y}_x$ obţinute, se poate trece la trasarea pe grafic a liniei de regresie teoretice (graficul 52).

Reprezentând liniile de regresie YY', XX' pe graficul 49, observăm cum concentrarea cercurilor reprezentând frecvenţele diferitelor caze se face vertical de-a lungul lui YY', orizontal de-a lungul lui XX'. Natural că punctul lor de intersecţie va constitui centrul de concentrare maximă. Acest aspect ne permite să pătrundem o nouă proprietate a liniilor de regresie.

Coeficienţii de regresie b_{12} şi b_{21} reprezintă pentru rezultatele unei atare cercetări ceva mai mult decât nişte simple valori numerice. Coeficientul b_{21}, de pildă, reprezintă variaţia mediei legate a lui y, corespunzătoare unei variaţii a lui

Scanned with OKEN Scanner

x egală cu unitatea de măsură. Astfel, în exemplul tratat, cum b_{21} este aproximativ $\frac{1}{2}$, înseamnă că unei creşteri a înălţimii părintelui egală cu 1 deget (sau 1 cm), îi va corespunde *în medie* o creştere de $\frac{1}{2}$ de deget (sau $\frac{1}{2}$ cm) pentru înălţimea descendenţilor.

TABELUL 113

Calculul liniei de regresie

$x-m_{10}$	- 8,19	- 6,19	- 4,19	- 2,19	- 0,19	+1,81	+3,81	+5,81	+7,81
$b_{21}(x-m_{10})$	- 4,34	- 3,28	- 2,22	- 1,16	- 0,10	+0,96	+2,02	+3,08	+4,14
$\tilde{y}_x$	64,32	65,38	66,44	67,50	68,56	69,62	70,68	71,74	72,80

Tabelul 108

139. În cazul când distribuţia celor două variabile se prezintă sub forma unei serii de valori individuale, calculul liniilor de regresie se sprijină pe aceleaşi formule de mai sus, este însă condus în alt chip.

Tabelul 114 ne arată cum trebuie desfăşurate calculele în cazul menţionat. Diferitele valori necesare calculării momentelor de ordinul întâi şi al doilea sunt obţinute făcând, după caz, suma algebrică a coloanelor α, β, α^2, β^2, $\alpha\beta$.

TABELUL 114

Calculul liniilor de regresie pentru cazul unei distribuţii de valori individuale

Nr.	α_i	β_i	α_i^2	β_i^2	$\alpha_i\beta_i$
(1)	(2)	(3)	(4)	(5)	(6)
1	+0,5	- 0,25	0,25	0,0625	- 0,125
2	- 1	- 8	1	64	+ 8
3	+6,25	+18	39,0625	324	+112,5
4	+1	+ 2,75	1	7,5625	+ 2,75
5	+4,75	+27,5	22,5625	736,25	+130,625
6	- 1	- 15	1	225	+ 15
7	- 1	+ 6	1	36	- 6
8	+6	+32,25	36	1040,0625	+193,50
9	- 1,5	- 2,75	2,25	7,5625	+ 4,125
10	+0,75	+10,75	0,5625	115,5625	+ 8,0625
11	- 4	- 13,5	16	182,25	+ 54
12	+4,75	+21,75	22,5625	473,0625	+103,3125
13	- 3,25	- 12	10,5625	144	+ 39
14	+3,5	+21,75	12,25	473,0625	+ 76,125
15	+1	- 1	1	1	- 1

Scanned with OKEN Scanner

(continuare tab. 114)

(1)	(2)	(3)	(4)	(5)	(6)
16	- 2	- 9	4	81	+ 18
17	+3,5	+20,5	12,25	420,25	+ 71,75
18	- 2,75	- 12	7,5625	144	+ 33
19	-	+11,75	-	138,0625	-
20	+2	+ 8	4	64	+ 16
21	-	- 4	-	16	-
22	- 2,25	- 4	5,0625	16	+ 9
23	- 1,5	- 6,25	2,25	39,0625	+ 9,375
24	+3,25	-	10,5625	-	-
25	- 4,75	- 8,5	22,5625	72,25	+ 40,375
26	+4,5	+19,5	20,25	380,25	+ 87,75
27	+1,25	- 3,25	1,5625	10,5625	- 4,0625
28	- 2	- 4	4	16	+ 8
29	+2,25	- 6	5,0625	36	- 13,5
30	+2	+ 7,5	4	56,25	+ 15
31	- 1,5	- 21,75	2,25	473,0625	+ 32,625
32	- 1,5	- 1,75	2,25	3,0625	+ 2,625
33	- 2,25	- 6,5	5,0625	42,25	+ 14,625
34	+3,25	+ 7,25	10,5625	52,5625	+ 23,5625
35	+0,5	- 8	0,25	64	- 4
36	-	+ 4	-	16	-
37	-	- 0,25	-	0,0625	-
38	+1,75	+ 1	3,0625	1	+ 1,75
39	+0,5	+ 2,5	0,25	6,25	+ 1,25
40	+0,5	- 8	0,25	64	- 4
41	-	+ 3,25	-	10,5625	-
42	- 6,5	- 17,5	42,25	306,25	+113,75
43	+4	+ 6,5	16	42,25	+ 26
44	+2,25	+12,25	6,25	150,0625	+ 30,625
45	- 2,25	- 0,25	5,0625	0,0625	+ 0,5625
46	+3	+17	9	289	+ 51
47	+3,5	+10	12,25	100	+ 35
48	+5	+17,5	25	306,25	+ 87,5
49	+2,5	+ 9	6,25	81	+ 22,5
50	- 2,75	- 15	7,5625	225	+ 41,25
51	+2,25	+ 6	5,0625	36	+ 13,5
52	+0,5	+16,25	0,25	264,0625	+ 8,125
53	+0,5	+ 8,75	0,25	76,5625	+ 4,375
54	- 5,25	- 19,5	27,5625	380,25	+102,375

Scanned with OKEN Scanner

(continuare tab. 114)

(1)	(2)	(3)	(4)	(5)	(6)
55	+5,5	+13	30,25	169	+ 71,5
56	- 1	+ 1,5	1	2,25	- 1,5
57	- 2	- 10	4	100	+ 20
58	+1,5	+ 3	2,25	9	+ 4,5
59	- 4	- 12,5	16	156,25	+ 50
60	- 2	- 5,25	4	27,5625	+ 10,5
61	- 2,25	- 9,75	5,0625	95,0625	+ 21,9375
62	- 2,5	+ 4,75	6,25	22,5625	- 11,875
63	+6,5	+23,5	42,25	552,25	+152,75
64	- 2,25	+ 1,25	5,0625	1,5625	- 2,8125
65	+5	+15	25	225	+ 75
66	- 2	- 14,5	4	210,25	+ 29
67	- 4	- 15	16	225	+ 60
68	-	+ 0,25	-	0,0625	-
69	+0,75	+ 3,25	0,5625	10,5625	+ 2,4375
70	- 3	- 24,5	9	600,25	+ 73,5
71	- 3	- 24,5	9	600,25	+ 73,5
72	- 3,25	- 19	10,5625	361	+ 61,75
73	-	+ 4	-	16	-
74	- 7,5	- 27,75	56,25	770,0625	+208,125
75	+1	+15,75	1	248,0625	+ 15,75
76	- 2,75	- 9	7,5625	81	+ 24,75

Tabelul 109

$$a = 59 \qquad m_{10}(a) = \frac{97{,}75 - 90{,}50}{76} = 0{,}095 \qquad m_{10} = 59{,}095$$

$$b = 85 \qquad m_{01}(b) = \frac{414{,}50 - 379{,}75}{76} = 0{,}457 \qquad m_{01} = 85{,}457$$

$$\mu_{20} = \frac{711{,}9375}{76} - 0{,}095^2 = 9{,}358574 \qquad \mu_{02} = \frac{12541{,}6875}{76} - 0{,}457^2 = 164{,}813355$$ [1]

$$\mu_{11} = \frac{2623{,}2500 - 48{,}8750}{76} - 0{,}095 \times 0{,}457 = 33{,}829940$$

$$b_{21} = 3{,}615 \qquad b_{12} = 0{,}205$$

140. *Regresie curbilinie. Dreapta celor mai mici pătrate.* Spuneam mai sus că, dintre toate tipurile de regresie, cel care se prezintă mai des în practică şi care este în acelaşi timp mai uşor de tratat este regresia liniară. Este foarte

[1] *Evident, fără corecţia Sheppard.*

Scanned with OKEN Scanner

important să relevăm, în legătură cu acest tip de regresie, faptul că, dacă regresia lui *y* față de *x* este liniară, nu urmează de aici că și regresia lui *x* față de *y* va fi tot liniară.

Să considerăm tabelul 115 și să reprezentăm grafic în raport cu un sistem de axe de coordonate cele două serii de medii legate statistice (graficul 53).

TABELUL 115

Contingența între numărul dungilor inelare și lungimea unui grup de heringi

Lungimea în cm (*y*)	Numărul dungilor inelare (*x*)										n_{ij}	$m_{10}^{(j)}$	$\tilde{x}_y$
	4	5	6	7	8	9	10	11	12	13			
26	-	1	-	-	-	-	-	-	-	-	1	5,00	3,6
27	-	3	-	-	-	-	-	-	-	-	3	,00	4,2
28	1	1	3	2	-	-	-	-	-	-	7	5,86	4,8
29	2	3	5	-	-	-	-	-	-	-	10	5,30	5,4
30	1	15	21	16	7	1	-	-	-	-	61	6,26	6,0
31	-	28	43	45	21	6	2	-	-	-	145	6,59	6,6
32	1	24	49	53	26	7	5	1	-	-	166	6,75	7,2
33	1	1	11	26	26	22	11	3	3	1	105	8,14	7,8
34	-	1	4	4	15	14	7	3	1	1	50	8,62	8,4
35	-	-	1	-	1	2	2	-	-	-	6	8,67	9,0
n_{ii}	6	77	137	146	96	52	27	7	4	2	554	-	-
m_{01}'	30,17	30,85	31,34	31,65	32,25	32,92	33,07	33,29	33,25	33,50	-	-	-
$\tilde{y}_x$	30,4	30,9	31,3	31,8	32,2	32,6	33,1	33,5	34,0	34,4	-	-	-

Datele reproduse după A. Bowley, ***op. cit.****, p. 504.*

$$a = 7 \qquad m_{10}(a) = \frac{32}{554} = 0{,}0578 \quad m_{10} = 7{,}0578$$

$$b = 31 \qquad m_{01}(b) = \frac{431}{554} = 0{,}7780 \quad m_{01} = 31{,}7780$$

$$m_{11}(a,b) = \frac{619}{554} = 1{,}1173 \quad \mu_{11} = 1{,}0723$$

$$\mu_{20} = \frac{1330}{554} - 0{,}0578^2 = 2{,}3974 \quad \mu_{02} = \frac{1369}{554} - 0{,}778^2 - 0{,}0833 = 1{,}7825.$$

Din acest grafic se vede clar că, dacă regresia lui *y* în *x* poate fi privită într-o oarecare măsură ca liniară, nu putem spune același lucru și despre regresia lui *x* în *y*.

Cu toate acestea, calculul lui μ_{11} se poate face întocmai ca mai sus, iar apoi se pot obține și valorile lui b_{12} și b_{21}, deci se pot determina niște linii drepte a căror ecuație este dată de formulele (99) și (99 bis). Dacă linia *YY'*, a cărei ecuație este:

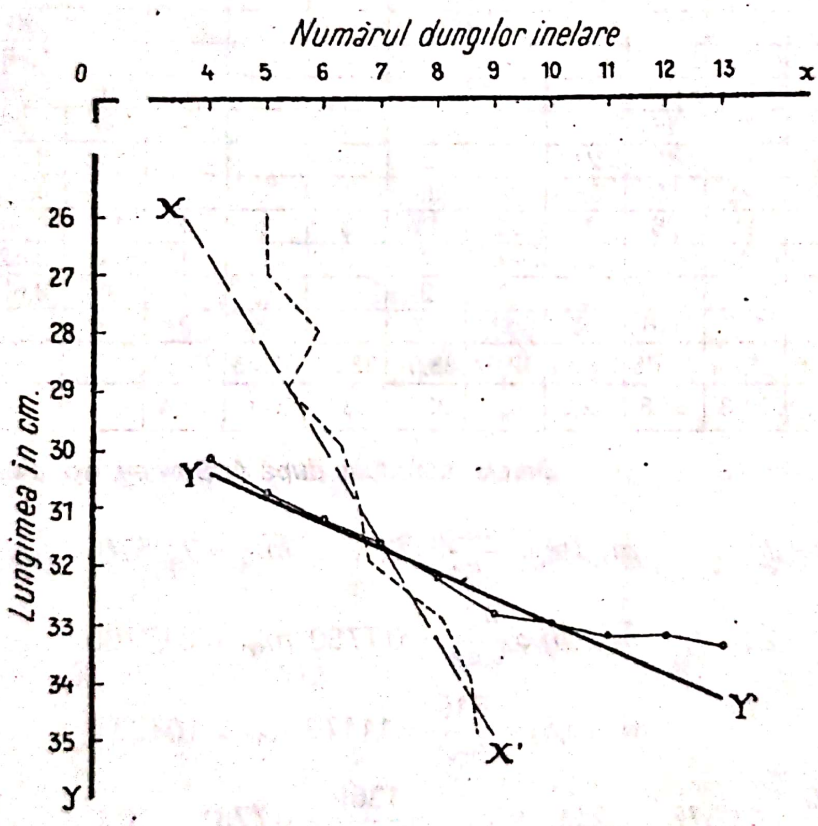
Scanned with OKEN Scanner

$$\tilde{y}_x = 31{,}778 + 0{,}447\ (x - 7{,}058)$$

poate fi considerată ca reprezentând regresia lui *y* în *x*, linia *XX'*:

$$X = 7{,}058 + 0{,}602\ (y - 31{,}778)$$

nu poate fi privită decât ca o reprezentare aproximativă a regresiei lui *x* în *y*. [1] O astfel de linie poartă numele de *drepta celor mai mici pătrate*. I se spune astfel pentru faptul că, dintre toate dreptele care se pot trage în planul celor două variabile, aceasta este cea pentru care suma pătratelor diferențelor $(Y - m_{01}^{(i)})$ sau $(X - m_{10}^{(j)})$ este cea mai mică cu putință. Dreapta celor mai mici pătrate fiind calculată după formula liniei de regresie, se confundă cu aceasta când regresia este liniară; astfel, această dreaptă n-are nici o legătură cu regresia și poate fi cel mult privită ca o aproximare a acesteia[2].



YY' = regresiunea lungimei în raport cu numărul dungilor
XX' = regresiunea numărului dungilor inelare în raport cu lungimea

Tabelul 115

Graficul 53. Regresiile dintre numărul dungilor inelare și lungimea heringilor.

[1] *Din această cauză am schimbat notația, punând X în locul lui* $\tilde{x}_y$.

[2] *Metoda pentru calculul liniei de regresie, în cazul când aceasta nu este liniară, este descrisă în* Apendice.

Scanned with OKEN Scanner

141. *Alt procedeu de calculare a coeficienţilor de regresie.* Momentul μ_{11} se mai poate calcula pe o cale diferită de cea expusă în paragraful 138 şi de care suntem nevoiţi să ne servim în anumite cazuri. Referindu-ne la tabelul 112 A, vom calcula suma produselor între frecvenţe şi abateri pentru fiecare coloană în parte. Aceste sume ne vor servi apoi la calculul lui $\sum n_{ij}\alpha_i\beta_j$.

Vom mai observa că termenii acestor sume conţin pentru aceeaşi coloană abaterea α_i ca factor, pe care, prin urmare, o putem scoate ca factor comun pentru a ţine seamă de ea după efectuarea sumei.

Astfel, pentru coloana corespunzătoare lui $x = 59{,}5$, vom avea $\alpha = -4$ şi, în consecinţă, suma produselor dintre abateri şi frecvenţe pentru această coloană va fi:

$$[(0{,}5) \times (-3) + 4 \times (-2) + 2 \times (-1)] \times (-4) = +46.$$

Suma din paranteză împărţită prin frecvenţa liberă a coloanei respective, 6,5, ne va da însă momentul de ordinul întâi al distribuţiei legate în raport cu originea $b = 68{,}5$. Prin urmare:

$$m_{11}(a,b) = \frac{\sum \alpha_i n_{i|} m_{01}^{(i)}(b)}{N}. \qquad (100)$$

Aplicând această formulă în cazul tabelului 108, vom obţine pe $m_{11}(a, b)$ pe calea indicată de tabelul 116.

TABELUL 116

Calculul momentului $m_{11}(a, b)$ prin formula 100

x	59,5	61,5	63,5	65,5	67,5	69,5	71,5	73,5	75,5
$n_{i\|} m_{01}^{(i)}(b)$	- 11,50	- 33,75	- 85,00	- 129,75	+ 7,00	+132,00	+141,00	+ 55,50	+ 9,00
$\alpha_i n_{i\|} m_{01}^{(i)}(b)$	+46,00	+101,25	+170,00	+129,75	-	+132,00	+282,00	+166,50	+ 36,00
$m_{01}^{(i)}(b)$	- 1,77	- 1,35	- 0,89	- 0,55	+ 0,02	+ 0,51	+ 1,11	+ 1,71	+ 1,64
$m_{01}^{(i)}$	64,96	65,80	66,72	67,40	68,54	69,52	70,72	71,92	71,78

Tabelul 108

Suma termenilor din linia a treia a acestui tabel este cea care constituie numărătorul lui $m_{11}(a, b)$ din relaţia (100) şi este evident egală cu valoarea obţinută mai sus: 1063,5. Cu ajutorul datelor din prima linie a aceluiaşi tabel se pot obţine momentele legate $m_{01}^{(i)}(b)$ şi deci mediile legate $m_{01}^{(i)}$.

Procedeul descris este singurul care poate fi aplicat pentru calculul liniei de regresie în cazul când legătura stocastică nu este cunoscută decât prin mediile legate statistice, cum este cazul tabelului 117.

Scanned with OKEN Scanner

TABELUL 117

Corespondența dintre înălțimea auriculară a capului și vârstă, pentru un grup de fete

Vârsta în ani (x)	Frecvențe n_{il}	Înălțimea medie a capului (mm) $m_{01}^{(i)}$	α_i	$m_{01}^{(i)}(b)$	$n_{il}\alpha_i m_{01}^{(i)}(b)$
(1)	(2)	(3)	(4)	(5)	(6)
3,5	1	115,25	-9	-8,75	+ 78,75
4,5	7	116,96	-8	-7,04	+ 394,24
5,5	18	117,47	-7	-6,53	+ 822,78
6,5	40	119,10	-6	-4,90	+ 1.176,00
7,5	76	120,30	-5	-3,70	+ 1.406,00
8,5	125	121,63	-4	-2,37	+ 1.185,00
9,5	177	121,72	-3	-2,28	+ 1.210,68
10,5	235	122,82	-2	-1,18	+ 554,60
11,5	261	123,14	-1	-0,86	+ 224,46
12,5	309	123,89	0	-0,11	-
13,5	263	124,86	+1	+0,86	+ 226,18
14,5	198	125,71	+2	+1,71	+ 677,16
15,5	214	126,16	+3	+2,16	+ 1.386,72
16,5	162	126,53	+4	+2,53	+ 1.639,44
17,5	95	126,91	+5	+2,91	+ 1.382,25
18,5	61	127,02	+6	+3,02	+ 1.105,32
19,5	13	129,56	+7	+5,56	+ 505,96
20,5	7	123,82	+8	- 0,18	- 10,08
21,5	8	126,50	+9	+2,50	+ 180,00
22,5	2	125,25	+10	+1,25	+ 25,00
	2.272	-	-	-	14.170,46

Datele reproduse din K. Pearson, ***On the General Theory of Skew Correlation and Non-linear Regression****, Drapers' Company Research Memoirs, Biometric Series II, p. 34.*

Alegând ca origini $a = 12,5$ ani și $b = 124$ mm, obținem în coloanele (4) și (5) abaterile α_i și momentele legate $m_{01}^{(i)}(b)$. Cu ajutorul acestor două coloane și prin procedeele cunoscute, obținem momentele:

$$m_{10}(a) = 0,2007 \qquad m_{01}(b) = 0,0469\ ^{1}.$$

[1] *Media momentelor legate de ordinul întâi este egală cu momentul liber de ordinul întâi.*

Scanned with OKEN Scanner

La fel :

$$m_{11}(a,b) = \frac{14180{,}54 - 10{,}08}{2272} = 6{,}2370$$

și deci:

$$m_{10} = 12{,}70 \qquad m_{01} = 124{,}05 \qquad \mu_{11} = 6{,}2276$$

Linia de regresie va avea ca ecuație:

$$\tilde{y}_x = 124{,}05 + 0{,}663(x - 12{,}70)\ ^{1}.$$

Tabele ca cel de mai sus se întâlnesc în special în cazurile în care nu interesează decât linia de regresie a lui y în x. Pentru astfel de distribuții, coeficientul de regresie b_{12} nu se poate stabili din cauză că valoarea lui μ_{02} nu poate fi determinată.

142. *Coeficientul de corelație pearsonian.* Dacă valoarea lui μ_{11} este egală cu zero, atunci ecuațiile liniilor de regresie devin:

$$\tilde{y}_x = m_{01} \qquad \tilde{x}_y = m_{10}$$

ceea ce înseamnă că mediile legate nu variază de la un caz la altul sau, cu alte cuvinte, că unei variații a uneia din variabile nu-i corespunde nici o variație a mediei legate în raport cu cealaltă variabilă. Acest fapt ne îndreptățește să privim astfel de variabile ca neavând între ele o legătură prea strânsă. Nu le vom putea numi independente, căci, cu excepția unui singur caz (§ 146), faptul că $\tilde{y}_x$ este constant nu înseamnă că și distribuțiile legate corespunzătoare diferitelor valori ale lui x nu vor varia împreună cu acesta. Ideea de lege stocastică este, după cum știm, mai cuprinzătoare decât simpla corespondență între $\tilde{y}_x$ și x.

Variabilele pentru care $\mu_{11} = 0$ poartă numele de *variabile necorelate*, termen introdus de prof. K. PEARSON.

Considerând acum cazul limită al unei legături rigide - când linia de regresie a lui y în x se confundă cu cea a lui x în y - și presupunând că regresia este liniară, vom avea:

$$b_{12} = \frac{1}{b_{21}}$$

sau:

$$\frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} = \pm 1\ . \qquad (101)$$

[1] *Aceasta trebuie să fie considerată mai degrabă ca dreapta celor mai mici pătrate, deoarece, în cazul de față, regresia pare să fie curbilinie (*K. Pearson, ***op. cit.**. p. 38).*

În oricare alt caz:

$$-1 < \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} < +1 \quad [1]. \tag{102}$$

Acest fapt l-a determinat pe prof. K. PEARSON, bazat în parte şi pe unele lucrări anterioare ale matematicianului BRAVAIS[2], să introducă celebrul său *coeficient de corelaţie* definit prin formula:

$$r = \frac{\mu_{11}}{\sqrt{\mu_{20}\mu_{02}}} = \frac{\mu_{11}}{\sigma_x \sigma_y} \quad [3]. \tag{103}$$

Acest coeficient de corelaţie ne indică prin valoarea lui o măsură a gradului de corelaţie între două variabile întâmplătoare. Într-adevăr, *el este egal cu 0 în cazul când variabilele sunt necorelate şi egal cu* ±1 *în cazul unei legături rigide liniare*[4].

Nu trebuie să uităm însă că *r* nu poate constitui o bună măsură a acestui grad de legătură decât în cazul când *regresiile sunt liniare.* Răspândirea profundă a întrebuinţării acestui coeficient în lucrările de statistică nu s-a făcut întotdeauna cu respectarea observaţiei de mai sus şi din acest punct de vedere faptul este de regretat. Pentru a se îngrădi întrebuinţarea nejustificată a lui *r*, vor trebui şi mai multe şi mai mari eforturi din partea metodologilor, decât au fost necesare pentru introducerea acestui mijloc de cercetare.

Şi pentru că problema este încă nepătrunsă de foarte mulţi cercetători, care apoi se miră în faţa rezultatelor absurde obţinute pe o cale condamnată de ştiinţă, voi da un exemplu de distribuţie în raport cu două variabile pentru care legătura este cvasirigidă, iar *r* este cu toate acestea nul (tabelul 118).

Acest tabel a fost astfel construit încât să pună în evidenţă cât se poate mai clar rezultatul absurd la care se poate ajunge printr-o nejustificată întrebuinţare a lui *r*. Este evident că, pentru tabelul 118, $\mu_{11} = 0$; variabilele respective sunt însă departe de a fi independente, căci $C = 0{,}82$.

---

[1] *Relaţia se poate stabili cu ajutorul identităţii lui Lagrange.*

[2] *A. Bravais,* ***Analyse mathématique sur les probabilités des erreurs de situation d'un point****, Acad. des Sciences, Mémoires présentés par divers savants, II-e série, IX, p. 255.*

[3] *Cu această nouă definiţie, coeficienţii de regresie se pot scrie:*

$$b_{21} = r\frac{\sigma_y}{\sigma_x} \qquad b_{12} = r\frac{\sigma_x}{\sigma_y}$$

forme care se întâlnesc destul de des.

[4] *Semnul lui r ne arată sensul variaţiilor simultane ale celor două variabile: r este pozitiv când variaţiile sunt de acelaşi sens şi negativ în caz contrar.*

Scanned with OKEN Scanner

TABELUL 118

Exemplu de variabile necorelate, dar cert asociate

A 2-a variabilă	Prima variabilă							
	1	2	3	4	5	6	7	Total
0	-	-	50	100	50	-	-	200
1	-	5	50	-	50	5	-	110
2	-	90	-	-	-	90	-	180
3	5	5	-	-	-	5	5	20
4	40	-	-	-	-	-	40	80
5	55	-	-	-	-	-	55	110
	100	100	100	100	100	100	100	700

Date fictive

143. *Cauzalitatea şi coeficientul de corelaţie.* Exemplele artificial construite care urmează ne vor permite să prezentăm coeficientul de corelaţie sub o formă mai sugestivă. Să considerăm distribuţiile din tabelele 119A şi 119B. Aceste tabele oglindesc rezultatele următoarei experienţe:

Un număr oarecare de zaruri sunt aruncate la întâmplare. Se notează pentru această primă aruncătură numărul zarurilor ce arată un număr par. În urmă, se lasă pe masă o parte din zaruri, însemnate în prealabil printr-o culoare deosebită, de pildă, iar restul zarurilor se aruncă din nou. Acestea, împreună cu zarurile păstrate, formează a doua aruncătură, pentru care se notează din nou frecvenţa zarurilor arătând un număr par. Obţinem pe această cale o corespondenţă de numere: frecvenţele zarurilor arătând un număr par din prima aruncătură şi din a doua aruncătură. Se repetă experienţa de un număr oarecare de ori, iar rezultatele parţiale se clasifică într-un tabel cu două intrări. Pe un astfel de tabel putem calcula coeficientul de corelaţie între cele două frecvenţe, stabilind astfel gradul de legătură dintre ele. Evident, valoarea lui r va fi cu atât mai mare cu cât raportul zarurilor păstrate de la o aruncătură la alta va fi mai ridicat. Pe de altă parte, este de aşteptat ca valoarea lui r să nu depindă decât de acest raport care este în definitiv o măsură a ansamblului de cauze care influenţează deopotrivă cele două fenomene. Ea va fi egală cu zero, în cazul când de la o aruncătură la alta nu se va păstra nici un zar, şi cu 1, când se păstrează toate zarurile.

Însemnând cu u numărul total al zarurilor şi cu v numărul celor ce se păstrează de la o aruncătură la alta, vom avea în mod teoretic:

$$r = \frac{v}{u} \,^{1}. \qquad (104)$$

[1] *Vezi G. Darmois,* ***Statistique mathématique****, Paris, Doin, 1928, p. 202.*

Prin urmare, coeficientul de corelație măsoară ceea ce este comun în sistemele de cauzalitate a două fenomene, cu condiția expresă ca acestea din urmă să poată fi reprezentate de o schemă ca cea de mai sus.

O experiență de acest gen pentru verificarea practică a relației 104 a fost efectuată în cadrul Seminarului de statistică matematică de la Școala de Statistică din București, de către dl C. IOSIF. Experiența a constat din aruncarea a 8 zaruri, din care s-a păstrat de la o aruncătură la alta: *a)* 2 zaruri, tabelul 119 A, *b)* 6 zaruri, tabelul 119 B.

TABELUL 119

Corelația dintre două aruncături consecutive de zaruri

A. (2 zaruri păstrate)

A doua aruncătură	Prima aruncătură									
Cazuri	0	1	2	3	4	5	6	7	8	Total
0	-	-	1	1	-	-	-			2
1	-	1	4	2	2	-	-	-		9
2	1	2	6	10	7	4	-	-	-	30
3	1	7	9	18	17	10	6	1	-	69
4	-	2	6	25	20	16	13	2	-	84
5	-	1	7	10	14	20	9	3	1	65
6	-	2	3	3	7	11	6	1	1	34
7		-	-	1	1	1	2	1	-	6
8			-	-	1	-	-	-	-	1
Total	2	15	36	70	69	62	36	8	2	300

B. (6 zaruri păstrate)

A doua aruncătură	Prima aruncătură									
Cazuri	0	1	2	3	4	5	6	7	8	Total
0	-	-	1							1
1	2	5	-	1						8
2	-	5	12	12	5					34
3		5	18	22	19	7				71
4			5	29	23	12	3			72
5				6	13	22	13	1		55
6					9	16	13	3	-	41
7						5	7	3	2	17
8							-	1	-	1
Total	2	15	36	70	69	62	36	8	2	300

Scanned with OKEN Scanner

Trebuie să relevăm faptul că, pentru anumite caze din aceste tabele, frecvența teoretică este nulă (aceste caze sunt barate în diagonală). Într-adevăr, în experiența corespunzătoare tabelului 119 A, dacă în prima aruncătură avem frecvența 0, evident că, din cauză că numai 6 zaruri sunt rearuncate, nu vom putea obține pentru cea de a doua aruncătură nici frecvența 7, nici 8.

Comparând acum, chiar fără să recurgem la o reprezentare grafică, tabelele între ele, vedem cum concentrația frecvențelor de-a lungul liniilor de regresie este mai mare în al doilea caz decât în primul. Este de așteptat, prin urmare, ca valorile coeficienților de corelație să nu prezinte același ordin de mărime.

Calculul coeficientului de corelație ne dă pentru aceste tabele:

Tabelul 119 A $r = 0{,}357$ (valoarea teoretică = 0,25)

Tabelul 119 B $r = 0{,}787$ (valoarea teoretică = 0,75)

144. *Abaterea-tip a coeficientului de corelație.* Diferențele între valorile statistice și cele teoretice constatate în paragraful precedent sunt datorate fluctuațiilor hazardului. Ele rămân însă în limita permisă acestor fluctuații. Într-adevăr, pentru cazul eșantioanelor numeroase, abaterea-tip a lui r este dată de formula:

$$\sigma_r = \frac{1-r^2}{\sqrt{N}}. \qquad (105)$$

Or, cum în cazul când N este mare, r se distribuie normal, îi putem aplica legea variației statistice.

Vom avea, prin aplicarea directă a formulei (105):

Tabelul 119 A $\sigma_r = \frac{1-0{,}25^2}{\sqrt{300}} = 0{,}054$

Tabelul 119 B $\sigma_r = \frac{1-0{,}75^2}{\sqrt{300}} = 0{,}025$

În nici unul din aceste cazuri diferența datorată hazardului nu întrece de două ori abaterea tip respectivă.

În exemplul considerat am putut face o aplicare strictă a formulei (105) pentru faptul că valoarea teoretică a lui r ne era cunoscută. În majoritatea cazurilor ce se prezintă în practică, nu putem cunoaște exact această din urmă valoare; o vom înlocui, așa cum am mai procedat și în alte cazuri, prin valoarea statistică respectivă, obținând astfel o evaluare destul de bună a lui σ_r.

Referindu-ne la tabelul 108, respectiv 112 B, vom avea:

Scanned with OKEN Scanner

$$r = \frac{\mu_{11}}{\sigma_x \sigma_y} = \frac{3{,}916}{7{,}443} = 0{,}526$$

de unde:

$$\sigma_r = \frac{1 - 0{,}526^2}{\sqrt{1078}} = 0{,}022$$

ceea ce ne îndreptăţeşte să scriem (§ 103):

$$r = 0{,}526 \pm 0{,}015 \ ^{1}.$$

145. *Coeficientul de corelaţie şi corecţia Sheppard.* Asupra calculului lui *r* mai sunt de spus câteva lucruri care privesc atât simplificarea operaţiilor, cât şi una din laturile cele mai delicate ale corecţiei Sheppard.

Trebuie să remarcăm, în primul rând, că pentru calculul lui *r* nu este nevoie să recurgem la medii, deoarece împărţirea prin numărul total al frecvenţelor se face atât la numitorul, cât şi la numărătorul formulei (103). De asemenea, nu mai este nevoie să ţinem seama de transformările unităţilor de măsură.

Făcând uz de notaţiile pe care le-am introdus în Capitolul V, § 76 şi punând:

$$s_{11}(a,b) = \sum n_{ij}(x_i - a)(y_i - b)$$

unde abaterile sunt măsurate în intervale, vom avea:

$$N\mu_{11} = \left[s_{11}(a,b) - \frac{s_{10}(a)s_{01}(b)}{N} \right] \times d_1 d_2$$

$$N\mu_{20} = \left[s_{20}(a) - \frac{\overline{s_{10}(a)^2}}{N} - \frac{N}{12} \right] \times d_1^2$$

$$N\mu_{02} = \left[s_{02}(b) - \frac{\overline{s_{01}(b)^2}}{N} - \frac{N}{12} \right] \times d_2^2$$

---

[1] *Aici este locul să menţionăm că această valoare verifică în parte legea eredităţii a lui Galton (F. Galton,* ***The Average of Each Several Ancestor to the Total Heritage of the Offspring,*** *Proceedings of the Royal Society of London, vol. LXI, 1897, p. 402). După această lege, corelaţia între caracterele părinţilor şi ale copiilor este egală cu ½. Dacă această valoare nu este cuprinsă între limitele variaţiei statistice ale valorii de mai sus se explică în primul rând prin faptul că legea lui Galton nu trebuie privită ca o lege matematică.*

Scanned with OKEN Scanner

și deci:

$$r = \frac{s_{11}(a,b) - \dfrac{s_{10}(a)s_{01}(b)}{N}}{\sqrt{\left[s_{20}(a) - \dfrac{\overline{s_{10}(a)^2}}{N} - \dfrac{N}{12}\right]\left[s_{02}(b) - \dfrac{\overline{s_{01}(b)^2}}{N} - \dfrac{N}{12}\right]}} \qquad (106)$$

Termenii $\frac{N}{12}$ corespund corecției Sheppard și trebuie lăsați deoparte, când nu avem de a face cu o grupare pe intervale.

Astfel, în cazul distribuției din tabelul 108 (112 A), avem:

$$s_{11}(a,b) - \frac{s_{10}(a)s_{01}(b)}{N} = 1063{,}5 - \frac{102{,}5 \times 84{,}5}{1078} = 1055{,}47$$

$$s_{20}(a) - \frac{\overline{s_{10}(a)^2}}{N} - \frac{N}{12} = 2092{,}5 - \frac{102{,}5^2}{1078} - \frac{1078}{12} = 1992{,}92$$

$$s_{02}(b) - \frac{\overline{s_{01}(b)^2}}{N} - \frac{N}{12} = 2118{,}5 - \frac{84{,}5^2}{1078} - \frac{1078}{12} = 2022{,}05$$

$$r = \frac{1055{,}47}{\sqrt{1992{,}92 \times 2022{,}05}} = 0{,}526$$

În ceea ce privește aplicarea uneia sau alteia din formule (103) sau (106), se ridică o chestiune care a fost prea puțin examinată până în prezent. Este vorba despre faptul că este cazul să se țină seamă de corecția Sheppard în calculul lui *r*. Majoritatea statisticienilor moderni opinează pentru o valoare a coeficientului de corelație calculată fără această corecție. Motivele pe care se sprijină această părere sunt în număr de două:

În primul rând, corecția Sheppard face ca numitorul lui *r* să se micșoreze, ceea ce poate conduce, în cazul unei corelații cvasirigide, la valori pentru *r* mai mari decât unitatea și astfel inegalitatea $|r| \leq 1$ n-ar mai fi satisfăcută.

Cel de al doilea motiv se bazează pe faptul că nu se cunoaște încă modul în care trebuie modificată legea variației statistice dacă se aplică pentru calculul unui moment corecția Sheppard. Cu alte cuvinte, chestiunile din cadrul problemei eșantioanelor n-au fost încă pe deplin examinate în raport cu introducerea acestei corecții.

Cu toate acestea, problema nu rămâne deloc pe un plan secundar. Într-adevăr, coeficientul de corelație al distribuției din tabelul 108 are ca valoare, în caz când nu se aplică termenul corectiv Sheppard, $r = 0{,}503$, valoare care,

Scanned with OKEN Scanner

spre deosebire de cea obținută mai sus, verifică legea eredității formulată de GALTON.

146. *Suprafața normală de corelație.* După cum am putut stabili, pornind de la forma curbei frecvențelor, mai multe categorii de distribuții în raport cu o singură variabilă, putem, pornind de la suprafața de corelație, să distingem diferite forme ale acesteia. Bineînțeles că, pentru cazul unei distribuții în raport cu două variabile, problema va fi mult mai complexă, suprafețele fiind mai greu de clasificat după formă în categorii bine circumscrise. Din această cauză, o astfel de clasificare nici nu s-a făcut încă și nici nu pare prea utilă. Există totuși un caz de distribuție care merită toată atenția și care prezintă un interes practic și teoretic în același timp. Este vorba despre tipul de distribuție normală, analog distribuției normale pentru o singură variabilă.

Dacă însemnăm cu z distanța de la un punct al suprafeței de corelație, obținută după modul arătat mai sus (§ 134), la punctul corespunzător din planul variabilelor, vom avea:

$$z = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}\, e^{-\frac{1}{2(1-r^2)}\left[\left(\frac{x-m_{10}}{\sigma_x}\right)^2+\left(\frac{y-m_{01}}{\sigma_y}\right)^2-2r\left(\frac{x-m_{10}}{\sigma_x}\right)\left(\frac{y-m_{01}}{\sigma_y}\right)\right]}. \qquad (107)$$

Această ecuație definește complet suprafața de corelație normală, noi însă avem nevoie să cunoaștem mai ales câteva aspecte intuitive ale acesteia.

Cea mai importantă caracteristică a distribuției normale în raport cu două variabile este faptul că toate distribuțiile întâlnite în tabelul de contingență corespunzător, fie legate, fie libere, sunt normale în sensul pe care acest cuvânt îl are pentru distribuțiile de o singură variabilă. Prin urmare, fiecărei valori a uneia din variabile îi va corespunde pe această suprafață o curbă, curba frecvențelor a distribuției legate respective, și care va avea forma *profilului* unui clopot. Suprafața întreagă va avea forma unui clopot deformat prin turtire (graficul 54).

Punctul cel mai înalt al suprafeței corespunde valorilor $x = m_{10}$, $y = m_{01}$. Recunoaștem aici o proprietate a curbei normale. Pornind de la acest punct, frecvențele (adică valorile lui z) descresc în toate direcțiile, descreșterea făcându-se întotdeauna după legea normală pentru o singură variabilă.

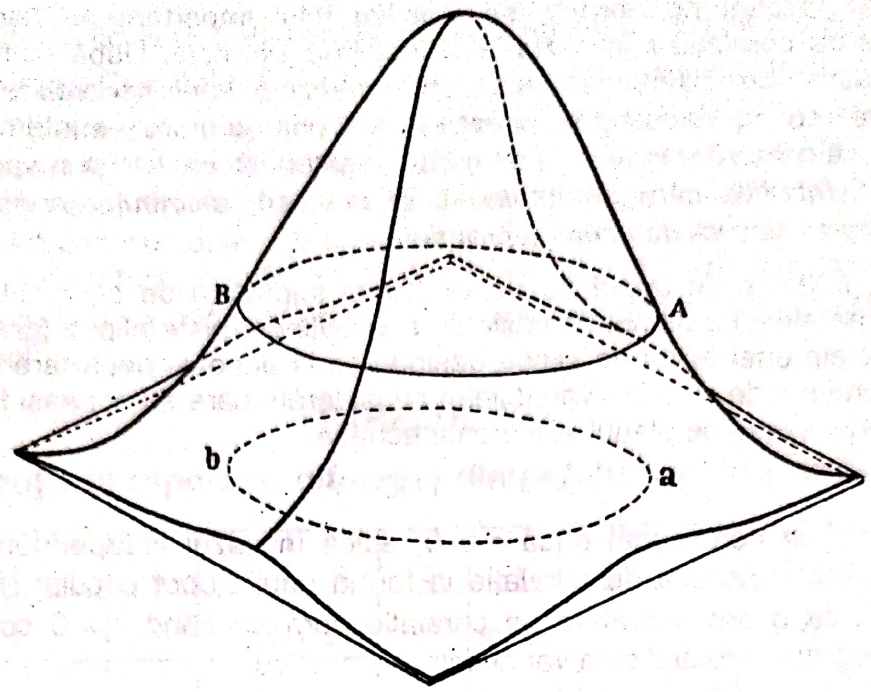
Însemnând prin X, Y abaterile relative:

$$X = \frac{x-m_{10}}{\sigma_x} \qquad Y = \frac{y-m_{01}}{\sigma_y}$$

ecuația suprafeței normale se poate scrie:

$$z = \frac{N}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}}\, e^{-\frac{1}{2(1-r^2)}(X^2+Y^2-2rXY)}. \qquad (107\text{ bis})$$

Scanned with OKEN Scanner



Graficul 54. Suprafaţa normală de corelaţie.

Distribuţiile libere ale celor două variabile considerate vor fi şi ele normale, aşa cum s-a spus mai sus. Ecuaţiile curbelor de frecvenţă respective vor fi, cu notaţiile adoptate:

$$\frac{N}{\sigma_x\sqrt{2\pi}}e^{-\frac{X^2}{2}} \qquad \frac{N}{\sigma_y\sqrt{2\pi}}e^{-\frac{Y^2}{2}}.$$

Revenind acum la ecuaţia (107 bis), observăm că dacă $r = 0$:

$$z = \frac{1}{N}\cdot\frac{N}{\sigma_x\sqrt{2\pi}}\cdot e^{-\frac{X^2}{2}}\cdot\frac{N}{\sigma_y\sqrt{2\pi}}\cdot e^{-\frac{Y^2}{2}}$$

adică:

$$n_{ij} = \frac{n_{i|}n_{|j}}{N}$$

relaţie care caracterizează completa independenţă a celor două variabile (90).[1]

[1] *Această demonstraţie ar fi trebuit să aibă ca suport volume infinitezimale ale suprafeţei de corelaţie. Făcută ca mai sus, pentru valorile cotei z, ea nu-şi pierde din valoare, este însă mai simplă pentru cel ce nu posedă calculul infinitezimal.*

Scanned with OKEN Scanner

Acest rezultat ne permite să stabilim rolul important pe care îl are coeficientul de corelaţie *r* în cazul unei distribuţii normale. După cum vedem, pentru o astfel de distribuţie, *faptul că variabilele sunt necorelate (r = 0) echivalează cu independenţa acestora* şi, prin urmare, suntem perfect îndreptăţiţi să măsurăm în acest caz gradul legăturii stocastice prin valoarea lui *r*. *Dintre distribuţiile care se întâlnesc în practică, distribuţia normală este singura care se bucură de această proprietate.*

147. *Elipsele de egală frecvenţă.* Tăind suprafaţa de corelaţie normală prin plane paralele cu planul variabilelor *x*, *y*, obţinem nişte elipse (graficul 54). Toate punctele unei astfel de elipse găsindu-se la aceeaşi depărtare de planul *x0y*, vor corespunde valorilor variabilelor considerate care au aceeaşi frecvenţă. Proiecţia unei elipse pe planul *x0y* are ca ecuaţie:

$$X^2 + Y^2 - 2rXY = c^2$$

şi nu devine un cerc decât dacă $r = 0$, adică în cazul independenţei dintre variabile, când suprafaţa de corelaţie ia forma unui clopot circular. Forma de clopot turtit ce o are suprafaţa de corelaţie normală când $r \neq 0$ corespunde existenţei legăturii stocastice a variabilelor respective[1].

148. *Distribuţia diferenţei a două variabile normal corelate.* O altă proprietate a distribuţiei normale în raport cu două variabile este faptul că suma sau diferenţa lor (în general, orice combinaţie liniară) se distribuie normal. Să ne referim la tabelul 108.

Lăsând deoparte faptul că distribuţia este făcută în raport cu o scară de intervale şi înlocuind pe fiecare din acestea printr-o valoare unică, putem deduce foarte uşor distribuţia cazurilor examinate în raport cu diferenţa dintre talia tatălui şi talia fiului (tabelul 120). Frecvenţa corespunzătoare unei diferenţe date se obţine, după cum este uşor de înţeles, adunând de la stânga spre dreapta în jos frecvenţele aceleiaşi diagonale.

Valorile caracteristice ale acestei distribuţii se pot deduce direct din cele obţinute deja pentru distribuţia în raport cu ambele variabile, în baza relaţiilor:

$$M = m_{10} - m_{01}$$

$$\sigma^2_{x-y} = \sigma^2_x + \sigma^2_y - 2r\sigma_x\sigma_y = \mu_{20} + \mu_{02} - 2\mu_{11} \qquad (108)$$

a căror verificare se poate face pornind chiar de la exemplul menţionat. O analiză mai amănunţită ne arată că ultima din egalităţile (108) nu mai poate fi adevărată dacă ţinem seamă şi de corecţiile Sheppard. Într-adevăr, dacă am calcula direct pe σ^2_{x-y}, trebuie să scădem corecţia $\frac{d^2}{12}$, în timp ce în membrul al

[1] *Pentru o confruntare în ceea ce priveşte elipsele de egală frecvenţă a rezultatelor date de practică cu cele indicate de teorie, a se vedea G. U. Yule,* **op. cit.**, *p. 326 - 328.*

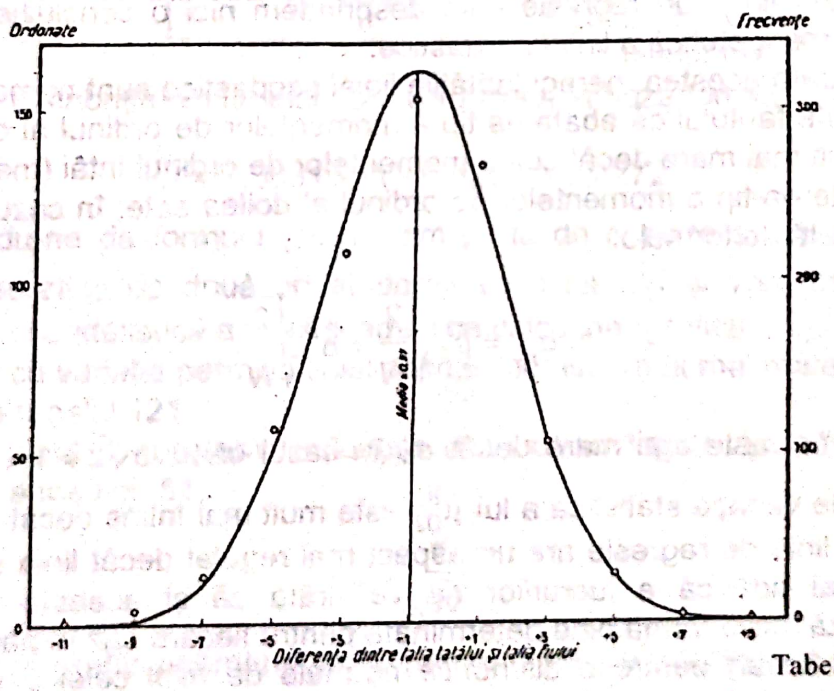
doilea, această corecție trebuie scăzută de două ori: o dată pentru μ_{20} și a doua oară pentru μ_{02}. Cea mai bună valoare a lui σ^2_{x-y} este cea calculată indirect, prin relația (108). Acest lucru se justifică prin faptul că gruparea pe intervale pentru o distribuție de felul celei din tabelul 120 se face în raport cu ambele variabile și că, prin urmare, intervalele unei astfel de distribuții se încalecă. Astfel, ultimul interval al cărui mijloc este 9 și ale cărui limite trebuie considerate egale cu (8, 10) se întinde de fapt de la 70,5 - 63,5 = 7, până la 72,5 - 61,5 = 11. Iată de ce este mai prudent să obținem pe σ_{x-y} cu ajutorul relației (108) decât să-l calculăm direct[1].

TABELUL 120

Distribuție în raport cu diferența dintre talia tatălui și talia fiului

Diferența în degete	-11	-9	-7	-5	-3	-1	+1	+3	+5	+7	+9
Frecvențe statistice	1	7,75	27,25	114,00	216,75	306,00	267,75	104,75	28,00	4,50	0,25
Frecvențe normale	0,3	4,1	27,2	105,7	239,3	315,9	243,2	109,0	28,6	4,3	0,4

Tabelul 108



Tabelul 120

Graficul 55. Distribuția valorilor diferenței dintre două variabile normal corelate.

[1] *Aceasta, în cazul variabilelor continue. Pentru cele discontinue, dificultatea de care ne ocupăm nici nu se prezintă.*

Scanned with OKEN Scanner

Vom avea, pentru tabelul 120:

$$M = 67{,}69 - 68{,}66 = -0{,}97$$

$$\sigma^2_{x-y} = 7{,}3948 + 7{,}5028 - 2 \times 3{,}1962 = 7{,}0652.$$

Normalitatea distribuției respective reiese atât din comparația directă a frecvențelor normale cu cele statistice, cât și din graficul 55.

149. *Linia scedastică*. Trecând acum mai departe la analiza variației distribuțiilor legate în raport cu variația variabilei respective, vom considera variația momentelor legate de ordinul al doilea, adică variația fluctuațiilor legate. Fiecărei valori x_i îi va corespunde o distribuție legată a cărei fluctuație o vom nota cu $\mu^{(i)}_{02}$. Corespondența între x și fluctuația legată se poate reprezenta grafic la fel ca și legătura dintre x și mediile legate. Curba corespunzătoare poartă numele de *linia scedastică* a lui y în x. Vom avea, de asemenea, o linie scedastică a lui x în y. Considerând din nou distribuția din tabelul 108 și calculând momentele $\mu^{(i)}_{02}$ pentru fiecare valoare a lui x, obținem corespondența din tabelul 121, reprezentată prin linia scedastică din graficul 56.[1]

Aspectul acestei curbe este din cele mai neregulate. Din el nu mai putem - ca în cazul liniei de regresie - să desprindem nici o concluzie în ceea ce privește forma ipotetică a liniei scedastice.

Cu toate acestea, neregularitățile liniei scedastice sunt numai aparente și sunt datorate faptului că abaterea tip a momentelor de ordinul al doilea este în general mult mai mare decât cea a momentelor de ordinul întâi (mediilor).

Abaterea-tip a momentelor de ordinul al doilea este, în cazul distribuțiilor normale, dată de formula:

$$\sigma_{\mu_2} = \mu_2 \sqrt{\frac{2}{N}} = \sigma^2 \sqrt{\frac{2}{N}} \qquad (109)$$

(tabelul 82) și este mai mare decât σ_M în cazul când $\sigma\sqrt{2} > 1$. Prin urmare, intervalul de variație statistică a lui $\mu^{(i)}_{02}$ este mult mai întins decât cel al mediei. Iată de ce linia de regresie are un aspect mai regulat decât linia scedastică. O analiză mai adâncă a lucrurilor ne va arăta că și aceasta din urmă se cristalizează într-o formă bine determinată pentru fiecare caz în parte.

Într-adevăr, pentru o distribuție normală de felul celei din tabelul 108, valoarea teoretică a fluctuației legate este:

$$\mu^{(j)}_{20} = \mu_{20}(1 - r^2) \qquad \mu^{(i)}_{02} = \mu_{02}(1 - r^2) \qquad (110)$$

[1] *Graficul este întocmit pe baza momentelor corectate pentru grupare.*

Aceasta înseamnă că fluctuaţia $\mu_{02}^{(i)}$ este aceeaşi pentru toate distribuţiile legate şi că, prin urmare, linia scedastică teoretică va fi o dreaptă paralelă cu axa absciselor[1].

TABELUL 121

Corespondenţa dintre fluctuaţiile legate şi valorile variabilei independente

x	$\mu_{02}^{(i)}$ (în intervale)	$\mu_{02}^{(i)}$ (corectat)	$n_i\mu_{02}^{(i)}$ (în intervale)
59,5	0,3286	0,9811	2,13590
61,5	0,8075	2,8967	20,18750
63,5	1,3605	5,1087	129,24750
65,5	1,5028	5,6779	356,91500
67,5	1,5948	6,0459	464,88420
69,5	1,3671	5,1351	352,02825
71,5	1,4175	5,3367	180,02250
73,5	1,8297	6,9855	59,46525
75,5	1,3104	4,9083	7,20720
			1572,09330

Tabelul 108

Pentru distribuţia din tabelul 108, vom avea, în baza formulei (110):

$$\mu_{02}^{(i)} = 7{,}5031 \times (1 - 0{,}526^2) = 5{,}4272.$$

Servindu-ne de formula (109), vom purta de o parte şi de alta a liniei scedastice teoretice de două ori abaterea tip a lui $\mu_{02}^{(i)}$ şi vom obţine astfel banda de variaţie statistică a liniei scedastice calculate (graficul 56). Vedem, pe această cale, că variaţia permisă liniei scedastice este mult mai mare decât cea prezentată de tabelul 121.

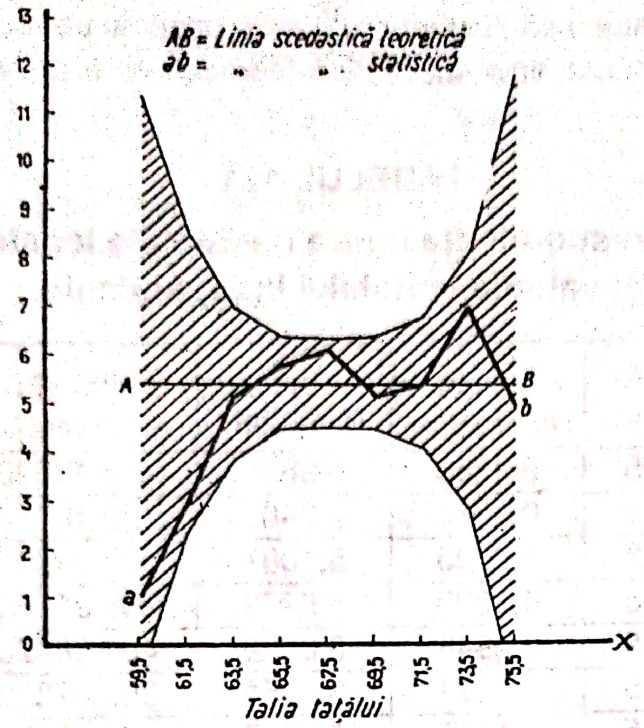
Acum, când cunoaştem fluctuaţiile distribuţiilor legate, utilizând formula abaterii medii a mediei (61):

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

putem face un grafic asemănător şi pentru linia de regresie a lui y în x (graficul 57).

---

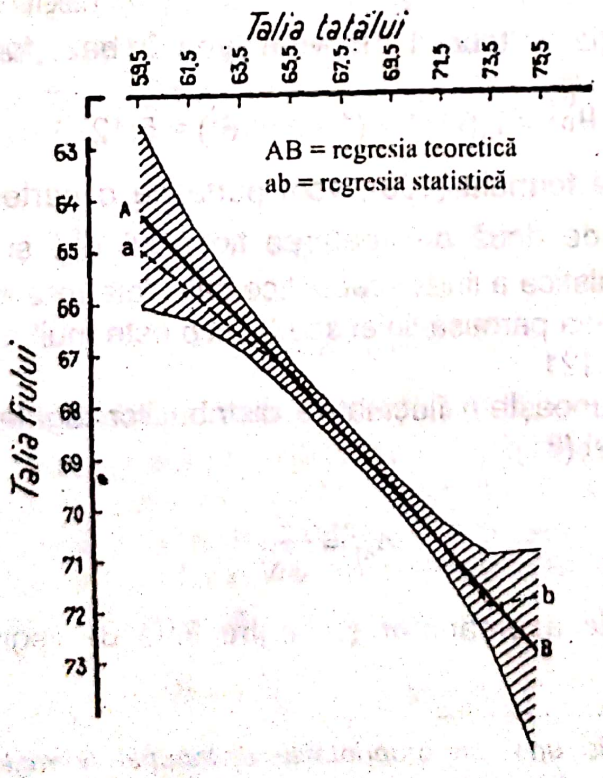
[1] *Acest lucru constituie una din proprietăţile distribuţiei normale, fără însă a fi o caracteristică a acesteia. Distribuţiile bucurându-se de constanţa fluctuaţiilor* $\mu_{02}^{(i)}$ *poartă numele de* homoscedastice *în x.*



Tabelul 121

Partea haşurată reprezintă regiunea de variaţie statistică.

Graficul 56. Comparaţie între liniile scedastice statistică şi teoretică.



Tabelul 113

Graficul 57. Regiunea de variaţie statistică a liniei de regresie.

Scanned with OKEN Scanner

Vom avea, pentru mediile legate:

$$\sigma_{m_{01}^{(i)}} = \sqrt{\frac{\mu_{02}^{(i)}}{n_{i|}}} = \sigma_y \sqrt{\frac{1-r^2}{n_{i|}}}. \qquad (111)$$

Din comparația graficelor 56 și 57 reiese și mai clar diferența din punct de vedere al stabilității formei dintre liniile scedastică și de regresie.

150. *Alt procedeu de calcul al coeficientului de corelație.* În cazul unei *distribuții homoscedastice și de regresie liniară,* relația (110) ne permite să obținem valoarea coeficientului de corelație pe o nouă cale:

$$r^2 = 1 - \frac{\mu_{02}^{(i)}}{\mu_{02}}. \qquad (112)$$

Formula aceasta are însă dezavantajul de a nu putea fi utilizată în practică din cauza abaterilor prea mari prezentate de valoarea statistică a lui $\mu_{02}^{(i)}$ față de valoarea teoretică corespunzătoare. Remediem acest lucru dacă înlocuim în (112) pe $\mu_{02}^{(i)}$ prin media aritmetică a tuturor valorilor similare, adică luând:

$$r^2 = 1 - \frac{1}{\mu_{02}} \cdot \frac{\sum n_{i|}\mu_{02}^{(i)}}{N} \qquad (112 \text{ bis})$$

unde, evident, vom atribui fiecărei valori $\mu_{02}^{(i)}$ o frecvență egală cu numărul elementelor cuprinse în distribuția legată corespunzătoare.

Relația (112 bis) mai prezintă și avantajul de a fi adevărată pentru *orice distribuție cu regresie liniară,* indiferent dacă este sau nu homoscedastică.

Pentru distribuția din tabelul 108, avem, servindu-ne de ultima coloană din tabelul 121:

$$\frac{\sum n_{i|}\mu_{02}^{(i)}}{N} = \frac{1572{,}0933}{1078} = 1{,}4583$$

și cum pentru aceeași distribuție:

$$\mu_{02} = 1{,}9591 \ ^{1}$$

vom avea:

---

[1] *Ambele valori sunt măsurate în intervale, transformarea în unități nefiind necesară pentru aceleași motive pe care le-am expus cu prilejul calculului lui r (§ 166).*

Scanned with OKEN Scanner

$$r^2 = 1 - \frac{1{,}4583}{1{,}9591} = 0{,}2556$$

deci:

$$r = 0{,}506.$$

Această valoare a fost obținută neținându-se seama de corecția Sheppard. Dacă vrem să obținem valoarea raportului de corelație, în urma aplicării acestei corecții, vom observa că, pentru $\mu_{02}^{(i)}$ corecția fiind aceeași $\left(\frac{1}{12}\right)$, aceasta se va transmite în întregime asupra mediei momentelor (§ 47) și, prin urmare:

$$r^2 = 1 - \frac{1{,}4583 - \frac{1}{12}}{1{,}9591 - \frac{1}{12}} = 0{,}2660$$

deci:

$$r = 0{,}516.$$

Ambele aceste valori diferă de cele obținute pe cale directă (§ 144, 145). Diferența provine din faptul că relația (112 bis) nu constituie decât o bună evaluare a lui r, egalitatea între valorile obținute pe ambele căi neproducându-se decât în cazul regresiei perfect liniare, condiție care nu poate fi îndeplinită decât de distribuțiile teoretice.

Dacă această condiție este îndeplinită și pentru regresia lui x în y, vom avea relația analoagă:

$$r^2 = 1 - \frac{1}{\mu_{20}} \cdot \frac{\sum n_{ij} \mu_{20}^{(j)}}{N} \qquad (112 \text{ ter})$$

Valorile lui r obținute din relațiile (112 bis) și (112 ter) sunt teoretic egale în cazul când ambele regresii sunt liniare, însă este de așteptat ca aceleași valori statistice să difere ușor între ele.

151. *Raportul de corelație.* Lăsând de o parte legătura ce există între coeficientul de corelație r și membrul al doilea al egalităților (112 bis) și (112 ter), prof. K. PEARSON a definit un nou instrument de măsură a legăturii stocastice dintre două variabile, prin relația:

$$\eta_{yx}^2 = 1 - \frac{1}{\mu_{02}} \cdot \frac{\sum n_{ij} \mu_{02}^{(i)}}{N} \qquad (113)$$

Scanned with OKEN Scanner

şi pe care l-a denumit *raportul de corelaţie a lui y faţă de x*. La fel se defineşte raportul de corelaţie a lui *x* faţă de *y*:

$$\eta_{xy}^2 = 1 - \frac{1}{\mu_{20}} \cdot \frac{\sum n_{ij} \mu_{20}^{(j)}}{N}. \qquad (113 \text{ bis})$$

Rapoartele de corelaţie, η_{xy} şi η_{yx}, se pot calcula *oricare ar fi natura distribuţiei considerate* şi au o interpretare precisă, independentă de cea a coeficientului de corelaţie, cu care nu devin teoretic egale decât în cazul regresiilor liniare.

Raportul de corelaţie este cel mai indicat instrument de măsură a gradului de legătură dintre două variabile. Într-adevăr, referindu-ne la relaţia (113), este clar că în cazul unei legături rigide neexistând decât o singură valoare a lui *y* pentru o valoare determinată a lui *x*, vom avea $\mu_{02}^{(i)} = 0$ şi deci $\eta_{yx} = 1$. Pe de altă parte, dacă există o perfectă independenţă între variabilele considerate, atunci nu este nici un motiv ca distribuţiile legate să se deosebească nici între ele şi nici faţă de distribuţia liberă corespunzătoare. Prin urmare, toate momentele acestor distribuţii vor fi egale între ele, deci $\mu_{02}^{(i)} = \mu_{02}$, de unde rezultă $\eta_{yx} = 0$.

Vedem, aşadar, că raportul de corelaţie are o interpretare precisă, oricare ar fi forma distribuţiei considerate, că variază între 0 şi 1, limite pe care nu le atinge respectiv decât în cazul independenţei absolute şi în cazul legăturii perfect rigide. În această ordine de idei, raportul de corelaţie constituie un instrument superior coeficientului de corelaţie care, atunci când este nul nu ne indică decât că variabilele sunt necorelate, iar pe de altă parte, în cazul unei legături rigide neliniare, poate avea o valoare diferită de 1.

Este de ajuns să menţionăm aici că, deşi valoarea coeficientului de corelaţie pentru distribuţia din tabelul 118 este egală cu zero, avem pentru aceeaşi distribuţie $\eta_{yx} = 0{,}97$, valoare care indică o legătură cvasirigidă. Rămânând pentru cazurile de regresie neliniară la coeficientul de corelaţie ca instrument de măsură a gradului de legătură stocastică, am fi, prin urmare, conduşi la concluzii false.

În ceea ce priveşte raportul lui η faţă de *r*, în cazul când regresia nu este liniară, se demonstrează că avem întotdeauna:

$$r \leq \eta \qquad (114)$$

egalitatea neavând loc decât dacă avem de-a face cu o regresie perfect liniară.

Calculul practic al raportului de corelaţie se bazează pe relaţia:

$$\eta_{yx}^2 = \frac{1}{\mu_{02}} \cdot \frac{\sum n_{ij} (m_{01}^{(i)} - m_{01})^2}{N}. \qquad (115)$$

Această relaţie se mai poate scrie:

Scanned with OKEN Scanner

$$\eta_{yx} = \frac{\sigma_{m_{01}^{(i)}}}{\sigma_y}$$

ceea ce înseamnă că η_{yx} este egal cu raportul dintre abaterea-tip a mediilor legate şi aceea a distribuţiei libere corespunzătoare, de unde şi termenul întrebuinţat pentru denumirea acestei măsuri a gradului de legătură stocastică.

Pentru aplicarea formulei (115) ne vom referi la distribuţia din tabelul 115. Calculul se conduce după modul arătat în tabelul 122.

TABELUL 122

Calculul raportului de corelaţie

$n_{i\|}$	$m_{01}^{(i)} - m_{01}$	$\left(m_{01}^{(i)} - m_{01}\right)^2$	$n_{i\|}\left(m_{01}^{(i)} - m_{01}\right)^2$
6	- 1,61	2,5921	15,5526
77	- 0,93	0,8649	66,5973
137	- 0,44	0,1936	26,5232
146	- 0,13	0,0169	2,4674
96	+0,47	0,2209	21,2064
52	+1,14	1,2996	67,5792
27	+1,29	1,6641	44,9307
7	+1,51	2,2801	15,9607
4	+1,47	2,1609	8,6436
2	+1,72	2,9584	5,9168
554	-	-	275,3779

Tabelul 115

$$\frac{\sum n_{i\|}\left(m_{01}^{(i)} - m_{01}\right)^2}{N} = \frac{275{,}3779}{554} = 0{,}4971$$

$$\eta_{yx} = \sqrt{\frac{0{,}4971}{1{,}8658}} = 0{,}516 \text{ (fără corecţia Sheppard)}$$

$$\eta_{yx} = \sqrt{\frac{0{,}4971}{1{,}7825}} = 0{,}528 \text{ (cu corecţia Sheppard)}$$

Pentru aceeaşi distribuţie:

$\eta_{xy} = 0{,}552$ $\quad r = 0{,}507$ (fără corecţia Sheppard)

$r = 0{,}519$ (cu corecţia Sheppard)[1]

[1] *Este clar că pentru η_{xy} nu este cazul să aplicăm această corecţie.*

Comparând valorile obţinute între ele şi rămânând, pentru a evita jocul produs de corecţia Sheppard, la valorile calculate fără această corecţie, vedem că diferenţa între η şi r este mai mare în cazul regresiei lui x în y, fapt explicabil dacă ţinem seamă că aceasta este departe de a fi liniară (§ 140) .

152. *Raportul de corelaţie şi corecţia Sheppard.* Aici este locul să revenim asupra efectului pe care introducerea corecţiei Sheppard îl are asupra valorilor coeficientului şi rapoartelor de corelaţie. Am văzut mai sus (§ 145) cum se pune problema pentru coeficientul r. În ceea ce priveşte raportul de corelaţie, trebuie să observăm, în primul rând, că identitatea între formulele (113) şi (115) nu are loc decât dacă fluctuaţiile nu cuprind termenul corectiv Sheppard. Nu trebuie, prin urmare, să fim surprinşi dacă, calculând raportul de corelaţie prin ambele formule, ţinând seama de corecţia Sheppard, vom obţine două valori diferite.

Astfel, în cazul distribuţiei din tabelul 108, obţinem, prin formula (115):

fără corecţia Sheppard $\eta_{yx} = 0{,}506$

cu corecţia Sheppard $\eta_{yx} = 0{,}517$

pe când ultimului caz îi corespunde, prin aplicarea formulei (113), valoarea 0,516 (§ 151). Se justifică astfel consideraţiile de mai sus.

În afară de acest fapt, introducerea corecţiei Sheppard în calculul lui η şi r poate să conducă la valori care să nu mai îndeplinească condiţia (114). Acesta este cazul distribuţiei din tabelul 108, pentru care am obţinut, cu corecţia Sheppard: $r = 0{,}526$ (§ 144), $\eta_{yx} = 0{,}517$ şi deci $r > \eta$. Nu se întâmplă acelaşi lucru pentru valorile calculate fără această corecţie, valori care satisfac întotdeauna inegalitatea menţionată.

După cum vedem, introducerea corecţiei Sheppard în calculul fie al lui r, fie al lui η, conduce la unele situaţii ciudate şi care n-au fost până în prezent lămurite. Iată de ce majoritatea statisticienilor sunt actualmente pentru o valoare în care nu intră termenii corectivi corespunzători.

153. *Criteriul Blakeman pentru liniaritatea regresiei.* Diferenţa $\eta^2 - r^2$ fiind întotdeauna pozitivă şi neanulându-se decât în cazul când regresia este pe deplin liniară, poate să fie considerată ca un criteriu de *liniaritate a regresiei*. Acest criteriu a fost propus de J. BLAKEMAN[1].

Evident că, din cauza fluctuaţiilor produse de hazard, diferenţa statistică $\eta^2 - r^2$ nu va fi nulă, chiar dacă regresia teoretică este liniară. Vom avea, prin urmare, nevoie să cunoaştem limita variaţiei statistice a acestei diferenţe.

Diferenţa $\eta^2 - r^2$ este considerată ca semnificativă dacă întrece cantitatea:

$$B = \frac{4}{\sqrt{N}}\sqrt{(\eta^2 - r^2)\left[(1-\eta^2)^2 - (1-r^2)^2 + 1\right]} \qquad (116)$$

[1] *J. Blakeman, On the Test for Linearity of Regression in Frequency Distributions, Biometrika, vol. IV, p.332.*

În cazul contrar, regresia corespunzătoare poate fi, în limitele fluctuațiilor produse de hazard, privită ca liniară.

Pentru distribuția din tabelul 115, avem:

$$\eta_{yx}^2 - r^2 = 0{,}0092 \qquad B = 0{,}016$$
$$\eta_{xy}^2 - r^2 = 0{,}0477 \qquad B = 0{,}036$$

Se confirmă astfel rezultatele la care ajunsesem deja pe alte căi mai puțin precise: regresia lui x în y nu este liniară.

154. *Considerații asupra coeficienților de măsură a gradului legăturii stocastice.* Pentru a încheia acest capitol, vom insista asupra unei chestiuni studiate deja în capitolul precedent (§ 132) și care ne va apărea acum sub o nouă lumină. Știm că pătratul mediu de contingență constituie o măsură absolută a depărtării unei distribuții de două variabile de distribuția independentă corespunzătoare. Este interesant de semnalat faptul că, în cazul unei distribuții normale, pentru care coeficientul de corelație r constituie și el o măsură a aceluiași fapt, există o relație simplă între acesta din urmă și pătratul mediu de contingență, φ^2. Această relație este:

$$r = \sqrt{\frac{\varphi^2}{1+\varphi^2}}. \qquad (117)$$

Făcând o comparație între (117) și formula care definește coeficientul de contingență (93), avem de relevat două lucruri.

În primul rând, formula (117) justifică forma adoptată de către prof. PEARSON pentru coeficientul de contingență, iar în al doilea, pune în evidență faptul că pentru distribuția normală $r = C$. Pentru o verificare a acestei proprietăți a distribuției normale, vom calcula pătratul mediu de contingență corespunzător tabelului 108; găsim:

$$\varphi^2 = 0{,}356266$$

deci:

$$C = 0{,}5125$$

valoare care nu diferă simțitor de cea a coeficientului de corelație obținută mai sus (§ 144). Vedem dar că, în cazul unei distribuții normale, toți coeficienții introduși până în prezent pentru măsura gradului de legătură stocastică între două variabile se confundă între ei și, prin urmare, teoretic cel puțin, este absolut indiferent la care din aceștia recurgem pentru a obține o măsură a corelației normale. În celelalte cazuri, numai considerarea în ansamblu a valorilor obținute pentru r, η și C ne poate oferi un aspect complet al chestiunii.

Valorile lui r și η nu pot fi însă obținute în cazul când distribuția este făcută în raport cu atribute nemăsurabile, cum este cazul celei din tabelul 101.

Scanned with OKEN Scanner

Dacă numai unul din atribute este măsurabil, atunci unul din rapoartele de corelație poate fi calculat; astfel, putem calcula pe η_{yx} dacă y este măsurabil, iar x nu.

Raportul de corelație și coeficientul de contingență nu pot fi, pe de altă parte, obținuți decât dacă distribuțiile se prezintă sub forma unui tabel de contingență. Pentru a calcula aceste valori în cazul unei distribuții în raport cu valori individuale, cum este cea din tabelul 109, suntem obligați să facem mai întâi o despuiere a datelor individuale în raport cu o scară de intervale pentru ca să obținem astfel tabelul de contingență corespunzător.

Nu este totuși de recomandat să aplicăm acești doi coeficienți eșantioanelor prea puțin numeroase, deoarece fluctuațiile datorate hazardului vor fi de un ordin de mărime care va afecta simțitor orice valoare dedusă dintr-un astfel de tabel de contingență. Tot în această ordine de idei, este bine ca pentru calculul lui C să nu întrebuințăm un tabel de contingență făcut în raport cu o scară de intervale prea fină, acest lucru ar conduce la frecvențe prea mici, precum și la o punere în evidență a excepțiilor specifice eșantionului corespunzător.

Scanned with OKEN Scanner

Capitolul XII

Alte aspecte ale corelației; corelație parțială

155. *Coeficientul de corelație tetrachorică.* Aplicațiile ideii de corelație în domenii din ce în ce mai variate și pentru cazuri care nu mai puteau fi tratate prin mijloacele expuse în capitolul precedent au condus, pe de o parte, la metode noi de calcul pentru coeficientul de corelație, iar pe de alta, la o analiză mai adâncă a legăturii stocastice.

Cazul unei 2×2 clasificări, care se întâlnește așa de des în aplicații și pentru care măsura gradului legăturii stocastice, prin metodele ce au făcut obiectul Capitolului X, întâmpină dificultățile menționate acolo, a atras de la început atenția cercetătorilor. Tot prof. K. PEARSON a fost cel care a imaginat două procedee pentru determinarea unor măsuri ale gradului de corelație pentru o astfel de clasificare.

Primul conduce la stabilirea unui coeficient analog coeficientului r și care a fost denumit *coeficientul de corelație tetrachorică.*

Pentru expunerea metodei vom lua ca suport distribuția din tabelul 108, pe care o vom prezenta însă sub forma unei 2×2 clasificări. În acest scop, vom grupa la un loc cazurile pentru care talia părintelui este sub 68,5 degete, iar cea a fiului sub 69,5, formând apoi cazele adiacente, așa cum se vede în tabelul 123, a cărui schemă generală este:

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	N

Prin intervertirea ordinii coloanelor și liniilor, ne putem întotdeauna aranja astfel ca: $a + c > b + d$ și $a + b > c + d$. Pentru un astfel de aranjament și în cazul unei distribuții normale, media primei variabile va cădea în prima coloană, iar media celei de-a doua, în prima linie. Ceea ce înseamnă că valorile care despart clasele tabelului se găsesc respectiv la dreapta și mai jos de mediile variabilelor considerate. Acest lucru se poate verifica imediat pe exemplul citat: $m_{10} = 67,69 < 68,5$, $m_{01} = 68,66 < 69,5$.

Scanned with OKEN Scanner

TABELUL 123

Corelația între talia părintelui și talia fiului

Talia fiului	Talia tatălui		
	sub 68,5	peste 68,5	Total
Sub 69,5	506	182	688
Peste 69,5	149,5	240,5	390
Total	655,5	422,5	1078

Tabelul 108

Presupunând acum că distribuția este normală, vom avea:

$$\frac{d}{N} = \frac{1}{2\pi\sqrt{1-r^2}} \int_h^\infty \int_k^\infty e^{-\frac{1}{2(1-r^2)}(x^2+y^2-2rxy)}\, dx\, dy \qquad (118)$$

unde h și k sunt abaterile relative ale valorilor care despart clasele distribuției. Determinarea acestora se face evident pornind de la relațiile:

$$\frac{b+d}{N} = \frac{1}{\sqrt{2\pi}} \int_h^\infty e^{-\frac{x^2}{2}}\, dx,$$

$$\frac{c+d}{N} = \frac{1}{\sqrt{2\pi}} \int_k^\infty e^{-\frac{y^2}{2}}\, dy. \qquad (119)$$

O dată valorile h și k determinate, $\mathbf{r}$ va fi obținut rezolvând ecuația (118). Pentru calculul efectiv al lui $\mathbf{r}$, aceasta din urmă este înlocuită prin ecuația echivalentă:

$$\frac{d}{N} = \tau_0(h)\tau_0(k) + \tau_1(h)\tau_1(k)\mathbf{r} + \tau_2(h)\tau_2(k)\mathbf{r}^2 + \ldots \qquad (118\text{ bis})$$

unde $\tau_i(h)$ este, după terminologia introdusă de PEARSON, *funcția tetrachorică de ordinul i.*[1]

Pentru aplicațiile practice nu este nevoie să cunoaștem mai mult în legătură cu aceste funcții, decât că pot fi obținute pentru diferite valori ale lui h (respectiv k) cu ajutorul unor table calculate de P.F. EVERITT (*Tables for Statisticians and Biometricians*, Part I, p. 42-51). Singurele funcții tetrachorice care se pot obține imediat sunt:

[1] *Vezi* Apendice.

$$\tau_0(h) = \frac{b+d}{N} \qquad \tau_0(k) = \frac{c+d}{N} \tag{120}$$

Referindu-ne la tabelul 123, vom avea (119):

$$\tau_0(h) = \frac{1}{\sqrt{2\pi}} \int_h^\infty e^{-\frac{x^2}{2}}\, dx = 0{,}39193,$$

$$\tau_0(k) = \frac{1}{\sqrt{2\pi}} \int_k^\infty e^{-\frac{y^2}{2}}\, dy = 0{,}36178.$$

Tablele menționate ne dau valorile celorlalte funcții tetrachorice, precum și valorile lui h în funcție de $\tau_0(h)$. Avem pe această cale și prin interpolare liniară:

$\tau_0(h) = 0{,}391$	$h = 0{,}27671$	$\tau_0(k) = 0{,}361$	$k = 0{,}35579$
$\tau_0(h) = 0{,}392$	$h = 0{,}27411$	$\tau_0(k) = 0{,}362$	$k = 0{,}35312$
$\tau_0(h) = 0{,}39193$	$h = 0{,}27429$	$\tau_0(k) = 0{,}36178$	$k = 0{,}35371$

Ca o verificare a calculelor efectuate până în prezent, vom compara valoarea lui h obținută mai sus cu cea pe care o putem deduce din tabelul de corelație din care derivă 2 × 2 clasificarea. Pentru aceasta din urmă avem:

$$h = \frac{68{,}5 - 67{,}69}{\sqrt{7{,}7284}} = 0{,}29137$$

valoare care nu diferă simțitor de precedenta.

Utilizând mai departe tablele, obținem:

$\tau_1(h) = 0{,}38421$	$\tau_1(k) = 0{,}37475$
$\tau_2(h) = 0{,}07452$	$\tau_2(k) = 0{,}09373$
$\tau_3(h) = -0{,}14506$	$\tau_3(k) = -0{,}13385$
$\tau_4(h) = -0{,}06292$	$\tau_4(k) = -0{,}07778$
$\tau_5(h) = 0{,}08958$	$\tau_5(k) = 0{,}07748$
$\tau_6(h) = 0{,}05598$	$\tau_6(k) = 0{,}06800$

Și, prin urmare, ecuația (118 bis) va fi:

$$-0{,}22310 + 0{,}14179 + 0{,}14398r + 0{,}00698r^2 + 0{,}01942r^3 + 0{,}00489r^4 + 0{,}00694r^5 + 0{,}00381r^6 = 0 \ ^{1} \tag{121}$$

---

[1] *Putem, și în unele cazuri suntem obligați chiar, să mergem mai departe decât gradul 6. Metoda prin care se obțin celelalte funcții tetrachorice este indicată în* ***Tables for Statisticians...****, loc. cit. A se vedea, de asemenea, Part II, p. 73-77.*

Scanned with OKEN Scanner

Vom căuta rădăcina acestei ecuaţii cuprinsă între 0 şi +1. În acest scop, vom întrebuinţa metoda de aproximaţie succesivă a lui Newton. Pentru început, vom adopta rădăcina ecuaţiei obţinute din precedenta prin suprimarea termenilor de grad superior lui 2. Obţinem **r** = 0,55, valoare care evident este mai mare decât cea care se obţine din consideraţia completă a ecuaţiei (121). Pentru aplicarea metodei lui Newton, vom însemna cu *F*(**r**) primul membru al acestei ecuaţii şi vom calcula:

$$F(0,55) = 0,00415.$$

Încercând cu **r** = 0,53 şi **r** = 0,52, avem:

$$F(0,53) = 0,00061 \qquad F(0,52) = -0,00112\ ^{1}.$$

Cum aceste două valori sunt de semne contrare, rădăcina căutată va fi cuprinsă între 0,52 şi 0,53. Admiţând că, atunci când intervalul nu este prea mare, creşterea lui *F*(**r**) este proporţională cu creşterea lui **r**, vom avea:

$$\mathbf{r} = 0,52 + 0,00112 \times 0,01/0,00173 = 0,526$$

care este valoarea coeficientului de corelaţie tetrachorică corespunzătoare tabelului 123. Egalitatea acesteia cu valoarea coeficientului de corelaţie obţinută pentru distribuţia iniţială (tabelul 108) vine să justifice pe cale de verificare procedeul expus în acest paragraf, cu toate că, în general, este de aşteptat o diferenţă între valorile obţinute pe aceste două căi.

156. *Cazul asociaţiilor de atribute calitative.* În ceea ce priveşte calculul corelaţiei tetrachorice, este remarcabil faptul că determinarea acesteia nu implică cunoaşterea valorilor pe care le iau variabilele în raport cu care este făcută distribuţia, ci numai a frecvenţelor din cele patru caze. Acest fapt ne face să ne gândim la aplicarea corelaţiei tetrachorice ca instrument de măsură şi pentru legătura stocastică dintre două variabile calitative, cum ar fi sexul, de pildă, sau dintre două variabile măsurabile exprimate însă prin calificative. Astfel, în cazul taliilor, am putea grupa indivizii pur şi simplu în scunzi şi înalţi.

Aplicarea corelaţiei tetrachorice pentru atare variabile se izbeşte însă de o dificultate de ordin teoretic: tot calculul acesteia se bazează pe faptul că distribuţia considerată este normală. Or, este greu de admis, cel puţin la prima vedere, ca o variabilă calitativă să dea naştere la o distribuţie normală, deci heterogradă. Din fericire, această dificultate este numai aparentă. Valorile multora din variabilele care nu se puteau măsura acum 100 de ani pot fi determinate azi cu o precizie uimitoare. Şi atunci de ce nu am admite că prin procedee care se vor descoperi în viitor vom putea înlocui calificativele de azi prin numere? Chestiunea distribuţiei normale a unei variabile care azi nu poate fi exprimată decât prin cuvinte, cum ar fi starea constituţiei fizice, dezvoltarea intelectuală, natura caracterului unui individ etc., nu mai trebuie privită ca un nonsens. Prin urmare, aplicarea coeficientului de corelaţie tetrachorică se poate extinde şi asupra acestor din urmă variabile.

---

[1] *Pentru a obţine mai uşor diferitele valori F(r) este recomandabil să ne servim de tablele puterilor numerelor întregi (**Tables for Statisticians...**, Part I, p. 38-41).*

Scanned with OKEN Scanner

Și este bine să notăm că tocmai acest lucru constituie importanța lui **r**, deoarece pentru variabilele măsurabile se poate oricând obține un tabel de contingență și calcula un coeficient de corelație obișnuit **r**.

Pentru exemplificarea acestor considerații, să considerăm distribuția din tabelul 97.

Avem în acest caz:

$\tau_0(h) = 0{,}378$ $\quad h = 0{,}31074$ $\quad \frac{d}{N} = 0{,}230$

$\tau_0(k) = 0{,}381$ $\quad k = 0{,}30286$

și:

$\tau_1(h) = 0{,}38014$	$\tau_1(k) = 0{,}38106$
$\tau_2(h) = 0{,}08353$	$\tau_2(k) = 0{,}08160$
$\tau_3(h) = -0{,}14021$	$\tau_3(k) = -0{,}14130$
$\tau_4(h) = -0{,}07001$	$\tau_4(k) = -0{,}06851$
$\tau_5(h) = 0{,}08432$	$\tau_5(k) = 0{,}08551$
$\tau_6(h) = 0{,}06182$	$\tau_6(k) = 0{,}06061$

Ecuația în **r** va fi:

$$F(\mathbf{r}) = -0{,}230 + 0{,}14402 + 0{,}14486\mathbf{r} + 0{,}00682\mathbf{r}^2 + 0{,}01981\mathbf{r}^3 + 0{,}00480\mathbf{r}^4 + 0{,}00721\mathbf{r}^5 + 0{,}00375\mathbf{r}^6 = 0$$

de unde, procedând ca mai sus:

$$\mathbf{r} = 0{,}550.$$

În fine, calculând corelația tetrachorică pentru distribuția din tabelul 96, obținem **r** = 0,188, valoare care ne conduce la aceleași rezultate ca cele obținute pe alte căi (§ 123, 129, 131).

157. *Coeficientul echiprobabil de corelație.* Singurul lucru pe care trebuie să îl avem în vedere atunci când aplicăm coeficientul tetrachoric este ca distribuția considerată să fie plauzibil normală. Pentru alte cazuri de distribuție nu avem altă modalitate decât determinarea coeficientului de satisfacție *P*, adică probabilitatea ca eșantionul considerat să provină dintr-o populație teoretică cu aceleași frecvențe marginale, însă pentru care atributele să fie independente (§ 131).

Dificultatea întrebuințării acestui din urmă procedeu constă din faptul că valorile lui *P*, pentru un grad de asociație evident, sunt de un ordin de mărime ce nu se poate exprima decât prin intermediul puterilor lui 10. PEARSON însuși spune că noi nu putem gândi în milioane de milioane. Acest lucru l-a determinat să construiască și pentru cazul distribuțiilor nenormale un coeficient de corelație, numit *coeficientul echiprobabil de corelație*[1]. Problema se pune în felul următor:

[1] *K. PEARSON, **On a Novel Method of Regarding Association of Two Variates Classed Solely In Alternate Categories**, Drapers' Company Research Memoirs, Biometric Series, VIII.*

Pornind de la o 2 × 2 clasificare, noi putem calcula probabilitatea P de care vorbeam mai sus. Cunoscând această probabilitate, vom determina valoarea coeficientului de corelație tetrachorică care ar avea probabilitatea P ca să provină dintr-o distribuție teoretică de atribute independente.

Această din urmă valoare, pe care o vom însemna prin r_P, este coeficientul echiprobabil de corelație. Termenul întrebuințat pentru denumirea lui oglindește ideea care stă la baza introducerii acestui coeficient.

Este necesar să relevăm faptul că r_P nu este un coeficient de corelație propriu-zis. El nu ne servește decât să transpunem pe o scară inteligibilă gradul de legătură pe care P îl măsoară printr-o variație aproape imposibil de urmărit.

Pentru a putea vorbi despre probabilitatea diferitelor valori statistice ale lui r, în cazul când pentru colectivitatea teoretică $r = 0$, trebuie să cunoaștem în primul rând abaterea-tip a coeficientului tetrachoric în aceste condiții. Însemnând-o prin ${}_0\sigma_r$, avem:

$$ {}_0\sigma_r = \frac{1}{\sqrt{N}HK}\sqrt{\frac{(a+b)(a+c)(d+b)(d+c)}{N^4}} \qquad (122) $$

unde:

$$ H = \frac{1}{\sqrt{2\pi}} e^{-\frac{h^2}{2}} \qquad K = \frac{1}{\sqrt{2\pi}} e^{-\frac{k^2}{2}} $$

h, k având semnificațiile de mai înainte.

Pentru calculul lui ${}_0\sigma_r$ ne putem servi de tablele special construite pentru acest scop (*Tables for Statisticians and Biometricians*, Part I, Table XXIV, p. 35).

Aceste table ne dau valorile lui:

$$ \chi_h = \frac{1}{H}\sqrt{\frac{(a+c)(b+d)}{N^2}} $$
$$ \chi_k = \frac{1}{K}\sqrt{\frac{(a+b)(c+d)}{N^2}} \qquad (123) $$

în funcție de $1 - \tau_0(h)$, $1 - \tau_0(k)$, adică de $\frac{a+c}{N}$, $\frac{a+b}{N}$. Astfel, pentru distribuția din tabelul 98:

$$ \frac{a+c}{N} = 0{,}620 \qquad \frac{a+b}{N} = 0{,}619 $$

și din table, prin interpolare liniară:

$$ \chi_h = 1{,}2744 \qquad \chi_k = 1{,}2756 $$

Scanned with OKEN Scanner

deci:

$$_0\sigma_r = \frac{1}{\sqrt{1000}} \times 1{,}2744 \times 1{,}2756 = 0{,}0514.$$

Pentru aceeași distribuție:

$$\chi^2 = 133{,}3265 \qquad \chi = 11{,}55$$

valoare care ne conduce pe calea cunoscută (§ 131) la:

$$\log P = -30{,}18518.$$

Cunoscând acum pe $_0\sigma_r$ și log P, cu ajutorul Tablei XVIII (*Tables for Statisticians*..., Part I, p. 31) se poate determina valoarea corelației tetrachorice statistice care are probabilitatea P ca să provină dintr-o colectivitate în care $\mathbf{r} = 0$. Din table obținem:

$$_0\sigma_r = 0{,}05 \begin{cases} r_P = 0{,}5 & -\log P = 26{,}025 \\ r_P = 0{,}6 & -\log P = 39{,}845 \end{cases}$$

$$_0\sigma_r = 0{,}06 \begin{cases} r_P = 0{,}6 & -\log P = 27{,}922 \\ r_P = 0{,}7 & -\log P = 41{,}634 \end{cases}$$

și, prin interpolare liniară în raport cu log P:

$$\left.\begin{array}{ll} _0\sigma_r = 0{,}05 & r_P = 0{,}5 + 0{,}1 \times \dfrac{4{,}160}{13{,}820} = 0{,}530 \\ _0\sigma_r = 0{,}06 & r_P = 0{,}6 + 0{,}1 \times \dfrac{2{,}263}{13{,}712} = 0{,}617 \end{array}\right\} -\log P = 30{,}185.$$

Interpolând în raport cu $_0\sigma_r$, obținem:

$$r_P = 0{,}530 + \frac{0{,}0014 \times 0{,}087}{0{,}01} = 0{,}542$$

care este coeficientul de corelație echiprobabilă al distribuției din tabelul 98.

Pentru aceeași distribuție am găsit mai sus $\mathbf{r} = 0{,}55$. Apropierea dintre această din urmă valoare și aceea a lui r_P are o interpretare depinzând de însăși esența coeficientului echiprobabil de corelație. Într-adevăr, distribuția considerată putând fi privită ca normală, măsura gradului legăturii stocastice dintre variabilele acesteia este legitim exprimată prin coeficientul de corelație tetrachorică. Or, cum r_P nu este decât o corelație echiprobabilă cu $\mathbf{r}$, în cazul considerat era de așteptat ca aceste două valori să nu difere simțitor între ele.

Scanned with OKEN Scanner

Concordanţa acestor două valori are însă loc numai în cazul când corelaţia tetrachorică are un sens şi când, prin urmare, calculul lui r_P devine inutil. Dacă însă corelaţia tetrachorică nu se poate aplica, distribuţia fiind departe de forma normală, valoarea lui r_P va fi singurul mijloc de măsură a asociaţiei şi va diferi de valoarea formală pe care am obţine-o pentru **r** printr-un calcul identic cu cel expus mai sus.

Să considerăm distribuţia din tabelul 100. Vom avea:

$$\frac{a+c}{N} = 0{,}9122 \qquad \frac{a+b}{N} = 0{,}9211$$

$$\chi_h = 1{,}7761 \qquad \chi_k = 1{,}8332$$

$$_0\sigma_r = \frac{1}{\sqrt{N}} \times \chi_h \times \chi_k = 0{,}0201$$

şi:

$$\chi^2 = 3256{,}7814 \qquad \chi = 57{,}07$$

de unde:

$$-\log P = 713{,}89769.$$

Vom determina pe r_P ca mai sus, servindu-ne de tablele indicate şi adoptând pentru simplificare:

$$_0\sigma_r = 0{,}02 \qquad \begin{matrix} r = 0{,}9 & -\log P = 556{,}100 \\ r = 0{,}95 & -\log P = 902{,}949 \end{matrix}$$

obţinem:

$$r_P = 0{,}9 + \frac{157{,}798 \times 0{,}05}{346{,}849} = 0{,}923.$$

Distribuţia considerată este departe de a fi normală şi acest fapt are ca urmare o diferenţă însemnată între r_P şi **r**, valoarea coeficientului de corelaţie tetrachorică fiind pentru aceeaşi distribuţie 0,652.

158. *Coeficientul biserial de corelaţie.* Se poate însă să fie nevoie de calculul unui coeficient de corelaţie pentru o distribuţie care s-ar prezenta sub forma tabelului 124, în care numai una din variabile este exprimată în 2 categorii alternative.

Soluţia acestei probleme a fost indicată tot de prof. K. PEARSON, definind *coeficientul biserial de corelaţie*[1]:

[1] *Karl Pearson, **On a New Method of Determining Correlation When One Variable is Given by Alternative and the Other by Multiple Categories**, Biometrika, vol. VII, p. 248-257.*

Scanned with OKEN Scanner

$$r_b = \frac{m_{10}^{(2)} - m_{10}}{\sigma_x} \times \frac{\tau_0(k)}{K}. \qquad (124)$$

Tabelul 124

Talia fiului	Talia tatălui									
	59,5	61,5	63,5	65,5	67,5	69,5	71,5	73,5	75,5	Total
Scund	6,5	25	87	190,5	197	135,25	40,25	5,5	1	688
Înalt	.	.	8	47	94,5	122,25	86,75	27	4,5	390
Total	6,5	25	95	237,5	291,5	257,5	127	32,5	5,5	1078

Tabelul 108

În această relație m_{10}, $m_{10}^{(2)}$ sunt mediile primei variabile, din care cea din urmă se referă la seria legată minoră, adică la aceea pentru care totalul frecvențelor este cel mai mic; $\tau(k)$ și K sunt funcțiile definite mai sus (119, 122). Pentru tabelul 124, avem:

$$m_{10} = 67{,}6902 \qquad m_{10}^{(2)} = \frac{27064}{390} = 69{,}3949 \qquad \sigma_x = 2{,}7799.$$

În ceea ce privește calculul lui $\frac{\tau_0(k)}{K}$, determinând întâi valoarea:

$$\tau_0(k) = \frac{390}{1078} = 0{,}3618$$

vom putea obține direct raportul căutat, prin intermediul Tablelor II din *Tables for Statisticians...*, Part II, p. 2-10[1]. Găsim în table:

$$\tau_0(k) = 0{,}361 \qquad \frac{\tau_0(k)}{K} = 0{,}96402$$

$$\tau_0(k) = 0{,}362 \qquad \frac{\tau_0(k)}{K} = 0{,}96577$$

și, prin interpolare liniară:

$$\tau_0(k) = 0{,}3618 \qquad \frac{\tau_0(k)}{K} = 0{,}96542$$

deci:

[1] *Există o diferență de notații în această privință, pe care o semnalăm pentru preîntâmpinarea oricărei dificultăți. În aceste table,* τ_0 *este însemnat prin* $\frac{1}{2}(1 - \alpha)$, *iar* K *(sau* H*) prin* z.

$$r_b = \frac{1,7047}{2,7799} \times 0,96542 = 0,592.$$

Dacă această valoare diferă oarecum de cea obținută pentru coeficientul de corelație dedus din tabelul inițial este pentru motivul că r_b are o abatere tip mai mare decât r și, prin urmare, câmpul de variație statistică a primului este mult mai larg.

În ceea ce privește aplicarea coeficientului biserial de corelație, nu trebuie neglijat faptul că el nu poate fi logic considerat ca o bună măsură a legăturii stocastice dintre două variabile, decât dacă acestea se distribuie normal.

159. *Aplicație.* Vom mai face încă o aplicație a coeficientului de corelație biserial. Să ne întoarcem la distribuția din tabelul 90. Am văzut (§ 120, 124) că cele două distribuții în raport cu circumferința craniană, pentru grupurile intelect bun și intelect slab, nu pot fi considerate ca provenind din aceeași colectivitate. Diferențierea gradului de dezvoltare a intelectului în raport cu circumferința craniană a fost, prin urmare, stabilită. Nu am stabilit însă gradul legăturii stocastice dintre aceste două variabile, lucru pe care îl vom face prin calcularea coeficientului de corelație biserial corespunzător distribuției considerate. Vom avea:

$$m_{10} = 56,4925 \qquad m_{10}^{(2)} = 56,0915 \qquad \sigma_x = 1,4423$$

$$\tau_0(k) = 0,1754 \qquad \frac{\tau_0(k)}{K} = 0,67945$$

deci:

$$r_b = 0,189$$

valoare care, interpretată în lumina rezultatelor deja obținute, ne arată că, deși existența unei legături stocastice între circumferința craniană și dezvoltarea intelectului nu poate fi negată, gradul acesteia este destul de slab. Iată de ce, simpla măsură a circumferinței craniene nu poate constitui un indice demn de încredere al gradului de inteligență. Nu tot așa ar fi stat lucrurile, dacă valoarea coeficientului de corelație ar fi fost categoric semnificativă.

160. *Asupra unui calcul.* În paragrafele precedente am avut foarte des ocazia să lucrăm cu valorile unei expresii de forma:

$$H = \frac{1}{\sqrt{2\pi}} e^{-\frac{h^2}{2}}$$

fără să le determinăm efectiv; tablele speciale la care am făcut apel ne-au permis să evităm acest lucru. Este util însă să cunoaștem și calea prin care am putea mânui această expresie fără intermediul tablelor speciale.

Scanned with OKEN Scanner

Să luăm, de pildă, cazul tratat în paragraful precedent: cunoscând valoarea $\tau_0(k) = 0{,}1754$, noi putem, prin intermediul tablelor de la sfârșitul acestei cărți, determina prin interpolare pe k:[1]

$$\tau_0(k) = 0{,}1762 \qquad k = 0{,}93$$

$$\tau_0(k) = 0{,}1736 \qquad k = 0{,}94$$

deci:

$$\tau_0(k) = 0{,}1754 \qquad k = 0{,}9331$$

Acum, pornind de la această valoare a lui k, putem determina pe K (notația din table este z):

$$k = 0{,}93 \qquad K = 0{,}2589$$

$$k = 0{,}94 \qquad K = 0{,}2565$$

de unde:

$$k = 0{,}9331 \qquad K = 0{,}2582$$

și, prin urmare:

$$\frac{\tau_0(k)}{K} = 0{,}6793$$

valoare care diferă de cea obținută prin tablele speciale. Diferența trebuie pusă atât pe seama operațiilor de interpolare liniară, care nu sunt decât aproximative, cât și pe seama faptului că de data aceasta am lucrat cu mai puține zecimale.

Se obținea aceeași valoare, dacă am fi făcut direct interpolarea lui K în raport cu valorile lui $\tau_0(k)$:

$$\tau_0(k) = 0{,}1762 \qquad K = 0{,}2589$$

$$\tau_0(k) = 0{,}1736 \qquad K = 0{,}2565$$

de unde, pentru $\tau_0(k) = 0{,}1754$:

$$K = 0{,}2565 + 0{,}0001 \times \frac{18 \times 24}{26} = 0{,}2582.$$

161. *Corelația parțială.* Să considerăm acum corelația dintre variabilele distribuției din tabelul 109: înălțimea și greutatea unui copil. Coeficientul de corelație este (§ 139):

[1] *Avem:* $\tau_0(k) = S(k) = 0{,}5 - S_1(k)$.

$$r = \frac{33{,}8299}{\sqrt{9{,}3586 \times 165{,}8134}} = 0{,}859.$$

Această valoare este însă departe de a oglindi gradul legăturii stocastice dintre cele două variabile, deoarece, în ultimă analiză, trebuie să ţinem seama de faptul că nu toţi copiii cercetaţi erau de aceeaşi vârstă, element care influenţează şi greutatea şi talia acestora.

Prin urmare, ceea ce noi luăm drept o corelaţie între aceste două elemente poate să nu fie decât o faţă a influenţei celui de-al treilea element, vârsta, asupra celor dintâi. Tot aşa, lipsa unei corelaţii dintre recolte şi temperatură poate să fie consecinţa influenţei pe care o au ploile asupra acestor variabile, pozitivă asupra primeia, negativă asupra celeilalte.

Vedem, aşadar, că, pentru o analiză mai adâncă a lucrurilor, coeficientul de corelaţie, în felul cum a fost definit în capitolul precedent, nu poate constitui decât o măsură a legăturii stocastice brute, adică aşa cum reiese din date. Este necesar, pentru completarea unui studiu în această direcţie, să ne fixăm în prealabil variabilele care pot interveni şi modifica gradul aparent al unei legături stocastice. Cel mai simplu mijloc prin care ne putem da seama dacă o corelaţie este numai aparentă sau dacă, prin jocul unui al treilea element, este redusă la o corelaţie nesemnificativă constă din comparaţia tuturor coeficienţilor obţinuţi prin gruparea variabilelor două câte două, în toate modurile posibile.

În cazul tabelului 109, însemnând înălţimea, greutatea, vârsta, respectiv prin indicii 1, 2, 3, vom avea pentru corelaţia deja calculată:

$$r_{12} = 0{,}859,$$

iar pentru corelaţiile dintre vârstă, pe de o parte, şi greutate şi înălţime, pe de alta:

$$r_{13} = 0{,}377 \qquad r_{23} = 0{,}381.$$

Faptul pe care-l bănuiam numai se confirmă acum: vârsta fiind în corelaţie pozitivă şi cu înălţimea şi cu greutatea, valoarea obţinută pentru r_{12} este influenţată de acest fapt şi, prin urmare, reprezintă un grad de legătură stocastică mai mare decât cel care există într-adevăr între înălţime şi greutate. Este clar că dacă observaţiile ar fi purtat asupra unor copii a căror vârstă ar fi variat nu între 13 şi 15 ani, ci între 2 şi 18 ani, de pildă, am fi obţinut valori şi mai ridicate şi pentru r_{12} şi pentru r_{13}, r_{23}.

Pentru a obţine o măsură a gradului adevărat de legătură între aceste două variabile, ar trebui, prin urmare, să pornim de la un grup de copii de *aceeaşi vârstă,* adică să eliminăm astfel influenţa celui de-al treilea element.

Recunoaştem aici unul din principiile călăuzitoare ale experienţei de laborator: menţinerea la acelaşi nivel a uneia sau a mai multor variabile care intervin în fenomenul studiat. Aplicarea acestui principiu în experienţele noastre este aproape o condiţie necesară pentru descoperirea legilor naturale. Numai că

Scanned with OKEN Scanner

posibilitatea aplicării lui efective este limitată la mijloacele de care dispunem. După cum spune prof. A. AFTALION, *la lumière des étoiles reste invisible le jour, s'il n'y avait jamais eu de nuit, nous ignorerons son existence; car pour connaître cette existence il faudrait pouvoir éteindre le soleil*[1].

Dacă, în exemplul de mai sus, izolarea influenței stocastice a vârstei asupra înălțimii și greutății se poate face prin observarea copiilor de aceeași vârstă numai, nu vom putea proceda la fel în toate cazurile. Instrumentul care ne va permite să obținem corelația adevărată dintre două variabile, chiar fără izolarea celorlalți factori care o pot influența, este *coeficientul de corelație parțială*, a cărui bază au pus-o F. Y. EDGEWORTH și K. PEARSON[2].

Coeficientul de corelație parțială între variabilele 1 și 2 se notează prin $r_{12\cdot34\cdots n}$ și măsoară corelația dintre acestea, influența celorlalte variabile, 3, 4, ..., n, fiind eliminată. Indicii ce urmează după punct poartă numele de indici secundari, iar numărul lor indică *gradul coeficientului de corelație parțială.* Astfel, $r_{12\cdot34}$ este un coeficient de gradul 2.

Spre deosebire de coeficienții de corelație parțială, coeficienților r_{12}, r_{13}, r_{23} care sunt de fapt de gradul 0, li s-a dat numele de *coeficienți de corelație totală.*

Trebuie considerat ca un câștig enorm pentru posibilitățile de cercetare a fenomenelor colective că s-a putut găsi mijlocul prin care se poate calcula un coeficient de corelație parțială, fără să fie nevoie de izolarea efectivă a variabilelor corespunzând indicilor secundari. Într-adevăr, orice coeficient de corelație parțială se poate calcula pornind de la indicii de corelație parțială de grad imediat inferior, prin intermediul formulei fundamentale:

$$r_{12\cdot34\cdots n} = \frac{r_{12\cdot34\cdots n-1} - r_{1n\cdot34\cdots n-1}r_{2n\cdot34\cdots n-1}}{\sqrt{1-r^2_{1n\cdot34\cdots n-1}}\sqrt{1-r^2_{2n\cdot34\cdots n-1}}}. \qquad (125)$$

Această formulă ne permite obținerea coeficienților de gradul 1, cu ajutorul celor de gradul 0, adică al coeficienților de corelație totală. Aceștia din urmă sunt singurii pe care îi putem deduce din datele statistice, ceilalți se obțin prin aplicarea succesivă a relației (125). În baza acesteia și utilizând valorile obținute pentru tabelul 109, avem:

$$r_{12\cdot3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r^2_{13}}\sqrt{1-r^2_{23}}} = 0{,}835$$

valoare care ne arată că gradul de corelație între greutate și înălțime, variațiile datorate vârstei fiind eliminate, este mai mic decât cel indicat de coeficientul de corelație totală.

---

[1] *A. Aftalion,* ***Cours de statistique****, Paris, Les Presses Universitaires de France, 1928, p. 224.*

[2] *F. Y. Edgeworth,* ***On Correlated Averages****, Philosophical Magazine, 1892, p. 194.*

K. Pearson, ***Regression, Heredity and Panmixia****, Philosophical Transactions, Series A, vol. CLXXXVII, 1896, p. 253.*

Scanned with OKEN Scanner

La fel:

$$r_{13\cdot 2} = \frac{r_{13} - r_{12}r_{32}}{\sqrt{1-r_{12}^2}\sqrt{1-r_{32}^2}} = 0{,}105 \qquad r_{23\cdot 1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}} = 0{,}120$$

ceea ce ne arată că între vârstă, pe de o parte, greutate și înălțime, pe de alta, nu există de fapt pentru copiii între 13 și 15 ani nici o corelație, valorile obținute fiind prea mici ca să le putem privi ca traducând o legătură reală. Înseamnă că, pentru perioada 13 până la 15 ani, creșterea copiilor se face în mod neregulat în ceea ce privește dezvoltarea în ansamblu a greutății și înălțimii.

Un alt exemplu instructiv pentru înțelegerea importanței coeficientului de corelație parțială ni-l oferă corelațiile dintre: (1) sănătatea mamei, (2) obiceiurile mamei și (3) mortalitatea copiilor (tabelul 125). Aplicând procedeul de calcul al corelației tetrachorice, avem:

$$r_{12} = 0{,}567 \qquad r_{13} = 0{,}329 \qquad r_{23} = 0{,}213.$$

Calculând acum corelația parțială dintre obiceiurile mamei și mortalitatea copiilor, găsim:

$$r_{23\cdot 1} = \frac{r_{23} - r_{21}r_{31}}{\sqrt{1-r_{21}^2}\sqrt{1-r_{31}^2}} = \frac{0{,}026457}{0{,}777856} = 0{,}034$$

ceea ce înseamnă că nu există nici o legătură stocastică directă între obiceiurile mamei și faptul dacă născutul moare sau trăiește.

TABELUL 125

Asociația dintre sănătatea mamei, obiceiurile ei și mortalitatea copiilor

Obiceiurile mamei	Sănătatea mamei		
	Bună	Slabă	Total
Bune	956	197	1153
Mediocre	257	286	543
Total	1213	483	1696

Copilul	Sănătatea mamei		
	Bună	Slabă	Total
În viață	1065	352	1417
Mort	148	131	279
Total	1213	483	1696

Scanned with OKEN Scanner

Copilul	Obiceiurile mamei		
	Bune	Mediocre	Total
În viață	997	420	1417
Mort	156	123	279
Total	1153	543	1696

Reprodus din W. Palin Elderton, ***Frequency Curves and Correlation,*** *London, Layton, 1927, p. 210.*

Obiceiurile mamei influențează asupra stării sănătății ei, care la rândul ei determină vitalitatea copilului. Astfel se explică de ce corelația totală între aceste două elemente este măsurată prin valoarea 0,213, de mică importanță, dar care totuși nu poate fi neglijată.

162. *Regresia multiplă.* Strâns legată de ideea de corelație este aceea de regresie care, ca și prima, poate fi extinsă unui ansamblu oarecare de variabile corelate. Există însă anumite rezerve în ceea ce privește generalizarea ideii de regresie. Știm că linia de regresie a unei variabile întâmplătoare în raport cu o alta este de fapt imaginea grafică a variației valorilor medii ale celei dintâi. Dacă pentru regresia dintre cele două variabile, cazul unei linii curbe de regresie se poate - deși mai greu - totuși trata, pentru un număr mai mare de variabile această problemă este intentabilă. De aceea, trecând la regresia multiplă, nu vom considera decât cazul regresiei liniare, care atunci când regresia efectivă nu îndeplinește această condiție, va trebui să fie privită sub unghiul principiului celor mai mici pătrate (§ 140).

Trecând acum la introducerea notațiilor necesare, vom însemna prin x_1, x_2, ..., x_n variabilele considerate, prin M_1, M_2, ..., M_n mediile aritmetice și prin σ_1, σ_2, ..., σ_n abaterile lor tip.

Dacă ne ocupăm numai de regresia dintre două variabile, a lui x_1 față de x_2, de pildă, soluția problemei ne este cunoscută; linia de regresie are ca ecuație (99):

$$x_{1\cdot 2} = M_1 - b_{12}M_2 + b_{12}x_2 \qquad (126)$$

unde $x_{1\cdot 2}$ înseamnă valoarea atribuită lui x_1 prin această relație. Valorile lui $x_{1\cdot 2}$ vor diferi în general de cele ale lui x_1, însă b_{12} este astfel determinat încât suma pătratelor diferențelor dintre aceste valori să fie cea mai mică cu putință. De fapt, aceste diferențe nu sunt decât niște abateri, abaterile valorilor statistice (x_1) față de valorile teoretice ($x_{1\cdot 2}$). Prin urmare:

$$\frac{\sum(x_{1\cdot 2} - x_1)^2}{N}$$

poate fi considerată ca o fluctuație. O vom însemna prin $\mu_{1\cdot 2}$, iar abaterea tip care se deduce de aici prin $\sigma_{1\cdot 2}$. Cu cât $\sigma_{1\cdot 2}$ va fi mai mic cu atât valorile teoretice furnizate de relația (126) vor interpreta mai exact realitatea statistică.

Scanned with OKEN Scanner

În ceea ce privește abaterea $\sigma_{1\cdot 2}$, ea este dată de relația:

$$\sigma_{1\cdot 2} = \sigma_1\sqrt{1-r_{12}^2} \tag{127}$$

formă care ne este deja cunoscută (110).

Pentru cazul general, vom avea:

$$x_{1\cdot 23\cdots n} = a + b_{12\cdot 34\cdots n}x_2 + b_{13\cdot 24\cdots n}x_3 + \ldots + b_{1n\cdot 23\cdots n-1}x_n \tag{128}$$

unde $x_{1\cdot 23\cdots n}$ este valoarea *teoretică* a variabilei x_1, dedusă din cunoașterea variabilelor x_2, x_3, ..., x_n. Coeficienții b poartă numele de coeficienți de regresie parțială și reprezintă partea din variația lui x_1 datorată variației unei singure variabile x. Observăm în ceea ce privește acești coeficienți că, spre deosebire de coeficienții de corelație parțială, ordinea primilor doi indici nu poate fi indiferentă, căci $b_{12\cdot 34\cdots n}$ diferă de $b_{21\cdot 34\cdots n}$, la fel ca și b_{12} de b_{21}, însăși semnificația lor fiind complet diferită.

Valoarea termenului a este dată de formula:

$$a = M_1 - b_{12\cdot 34\cdots n}M_2 - b_{13\cdot 24\cdots n}M_3 - \ldots b_{1n\cdot 23\cdots n-1}M_n. \tag{129}$$

Rămâne acum să determinăm coeficienții b. Aceștia se pot obține printr-o operație de iterare, analoagă celei pentru coeficienții de corelație parțială:

$$b_{12\cdot 34\cdots n} = \frac{b_{12\cdot 34\cdots n-1} - b_{1n\cdot 34\cdots n-1}b_{n2\cdot 34\cdots n-1}}{1 - b_{2n\cdot 34\cdots n-1}b_{n2\cdot 34\cdots n-1}} \tag{130}$$

și care ne permite obținerea oricărui coeficient de regresie parțială, imediat ce cunoaștem coeficienții de gradul imediat inferior. Această formulă are dezavantajul de ordin practic că, din cauză că $b_{12\cdot n}$ nu este același cu $b_{21\cdot n}$, coeficienții de regresie parțială vor fi de două ori mai numeroși decât cei de corelație parțială de același grad și, prin urmare, vom fi nevoiți să facem un parcurs mai lung pentru a obține valoarea unui b decât pentru a calcula coeficientul de corelație corespunzător. Pe când $r_{12\cdot 3}$ nu necesită decât cunoașterea valorilor r_{12}, r_{13}, r_{23}, pentru $b_{12\cdot 3}$ vom avea nevoie să determinăm în prealabil pe b_{12}, b_{13}, b_{32}, b_{23}. Vom ajunge imediat la formule mai nimerite pentru aplicațiile practice.

163. *Corelația multiplă.* Abaterea tip a lui $x_{1\cdot 23\cdots n}$ va fi dată de o formulă analoagă cu (127):

$$\sigma_{1\cdot 23\cdots n}^2 = \sigma_{1\cdot 23\cdots n-1}^2(1 - r_{1n\cdot 23\cdots n-1}^2). \tag{127 bis}$$

Dacă în această relație înlocuim pe $\sigma_{1\cdot 23\cdots n-1}^2$ în funcție de $\sigma_{1\cdot 23\cdots n-2}^2$, apoi pe acesta din urmă în funcție de $\sigma_{1\cdot 23\cdots n-3}^2$ etc., vom obține în definitiv:

Scanned with OKEN Scanner

$$\sigma^2_{1\cdot 23\cdots n} = \sigma^2_1 (1 - r^2_{12})(1 - r^2_{13\cdot 2})(1 - r^2_{14\cdot 23}) \cdots (1 - r_{1n\cdot 23\cdots n-1}) \qquad (131)$$

relație mai ușor întrebuințată în practică.

Determinarea lui $\sigma^2_{1\cdot 23\cdots n}$ ne interesează din două puncte de vedere: în primul rând, pentru cunoașterea aproximației obținută prin $x_{1\cdot 23\cdots n}$, iar în al doilea rând, pentru determinarea *coeficientului de corelație multiplă*. Pentru definirea acestui din urmă coeficient, să observăm că din (127) scoatem:

$$r^2_{12} = 1 - \frac{\sigma^2_{1\cdot 2}}{\sigma^2_1}$$

ceea ce înseamnă că, pornind de la $\sigma_{1\cdot 2}$ și σ_1, putem determina corelația dintre variabilele x_1 și x_2. Imitând acest procedeu, vom defini corelația lui x_1 față de ansamblul (x_2, x_3, ... x_n) prin relația:

$$R^2_{1\cdot 23\cdots n} = 1 - \frac{\sigma^2_{1\cdot 23\cdots n}}{\sigma^2_1}. \qquad (132)$$

$R_{1\cdot 23\cdots n}$ este coeficientul de corelație multiplă și măsoară gradul de legătură stocastică dintre x_1 și valoarea corespunzătoare $x_{1\cdot 23\cdots n}$ obținută din ecuația de regresie. R se poate exprima numai în funcție de corelațiile totale. Astfel, pentru cazul a trei variabile, vom avea:

$$\frac{\sigma^2_{1\cdot 23}}{\sigma^2_1} = (1 - r^2_{12})(1 - r^2_{13\cdot 2}) = \frac{1 - r^2_{12} - r^2_{23} - r^2_{31} + 2r_{12}r_{23}r_{31}}{1 - r^2_{23}} \qquad (133)$$

și, prin urmare:

$$R^2_{1\cdot 23} = \frac{r^2_{12} + r^2_{13} - 2r_{12}r_{23}r_{31}}{1 - r^2_{23}}. \qquad (132\text{ bis})$$

În toate celelalte cazuri este mai indicat, pentru aplicații numerice, să calculăm întâi pe $\sigma_{1\cdot 23\cdots n}$ prin intermediul relației (131) și apoi să obținem pe R.

Faptul că pentru determinarea valorilor variabilei x_1 prin regresia multiplă ținem seama nu numai de o singură variabilă cu care ea este corelată, ci de un ansamblu de astfel de valori, înseamnă o restrângere a limitelor variației statistice a lui x_1 și, prin urmare, o mărire a gradului corelației între valorile statistice și cele calculate.

Aceasta înseamnă că $R_{1\cdot 2\cdots n}$ va fi mai mare decât oricare alt coeficient de corelație totală dintre variabila x_1 și oricare din variabilele x_2, x_3, ..., x_n.

Scanned with OKEN Scanner

164. *Mărimea corelației multiple.* Dându-se un sistem de trei variabile x_1, x_2, x_3, putem calcula trei coeficienți de corelație multiplă, și anume $R_{1\cdot 23}$, $R_{2\cdot 31}$, $R_{3\cdot 12}$. Valorile acestora depind de valorile rapoartelor $\frac{\sigma_{1\cdot 23}}{\sigma_1}$, $\frac{\sigma_{2\cdot 31}}{\sigma_2}$, $\frac{\sigma_{3\cdot 12}}{\sigma_3}$ și vor fi cu atât mai mari cu cât acestea din urmă vor fi mai mici. Relația (133) ne arată însă că, sub această formă, numărătorul rapoartelor considerate este același pentru oricare din ele, deci diferența de la unul la altul este determinată numai de numitor. De aici, ținând seama de relația (132), deducem că cel mai mare coeficient de corelație multiplă ce se poate obține pornind de la un sistem dat este acela care corespunde unei regresii în raport cu cele mai slab corelate dintre variabile: $R_{1\cdot 23}$ va fi cel mai mare dintre coeficienții de corelație multiplă dacă r_{23} este cel mai mic dintre coeficienții de corelație totală[1]. Referindu-ne la distribuția din tabelul 109, vom avea:

$$R_{1\cdot 23} = 0{,}8616 \quad R_{2\cdot 13} = 0{,}8622 \quad R_{3\cdot 12} = 0{,}3931.$$

Prin aceste valori verificăm în primul rând faptul că valoarea oricăruia din acești coeficienți este mai mare decât valorile coeficienților de corelație totală dintre variabilele corespunzătoare, adică, de pildă:

$$R_{3\cdot 21} > r_{32} \quad R_{3\cdot 21} > r_{31}$$

iar în al doilea, constatăm că ordinea de mărime a coeficienților R este inversă celei a coeficienților de corelație totală dintre variabilele corespunzătoare indicilor secundari:

$$R_{2\cdot 13} > R_{1\cdot 32} > R_{3\cdot 21}$$

pe când:

$$r_{13} < r_{32} < r_{21}\,.$$

Din considerarea acestor inegalități desprindem în primul rând constatarea că vom avea mai mult succes în determinarea greutății pornind de la talie și vârstă decât în calcularea acesteia din urmă cunoscând talia și greutatea. Pe de altă parte, vedem că valorile lui $R_{1\cdot 23}$ și $R_{2\cdot 31}$ nediferind aproape deloc de cea a lui r_{12}, determinarea greutății pornind de la talie sau a taliei pornind de la greutate va fi prea puțin ameliorată în precizie dacă vom ține seama și de vârstă.

Acest din urmă rezultat nu trebuie să fie greșit interpretat. Într-adevăr, cunoașterea vârstei nu constituie un element hotărâtor pentru determinarea taliei sau greutății, însă acest lucru nu este valabil decât aplicat colectivității din studiul căreia am desprins rezultatele de mai sus, adică numai copiilor între 13

[1] *Acest rezultat se poate extinde și pentru cazul mai multor variabile, forma pe care o îmbracă nu va mai fi însă tot așa de simplă ca cea de mai sus.*

Scanned with OKEN Scanner

și 15 ani. Nu am fi ajuns la aceleași concluzii dacă vârsta copiilor examinați ar fi prezentat o variație mai mare, căci este știut că pentru copii și talia și greutatea au o variație specifică în raport cu vârsta. Așa că rezultatul obținut nu poate fi aplicat decât unui anumit grup pentru care vârsta prezintă variații prea mici ca să fie luate în seamă.

Exemplul examinat nu ar fi putut fi mai nimerit ales ca să constituie un avertisment contra intențiilor de a aplica rezultatele obținute din studiul unui simplu grup la cazul unei clase mai cuprinzătoare.

165. *Calculul coeficienților de regresie parțială.* O nouă formulă a coeficienților de regresie parțială este următoarea:

$$b_{12\cdot 34\cdots n} = r_{12\cdot 34\cdots n}\frac{\sigma_{1\cdot 34\cdots n}}{\sigma_{2\cdot 34\cdots n}} \tag{134}$$

relație care, pentru cazul a două variabile, conduce la formula cunoscută (§ 142):

$$b_{12} = r_{12}\frac{\sigma_1}{\sigma_2}.$$

Înlocuind în (134) pe r și σ prin relațiile (125) și (127), obținem:

$$b_{12\cdot 34\cdots n} = \frac{r_{12\cdot 34\cdots n-1} - r_{1n\cdot 34\cdots n-1}r_{2n\cdot 34\cdots n-1}}{1-r^2_{2n\cdot 34\cdots n-1}}\cdot\frac{\sigma_{1\cdot 34\cdots n-1}}{\sigma_{2\cdot 34\cdots n-1}} \tag{135}$$

formulă comodă în special pentru aplicațiile de ordin practic.

Vom scrie acum explicit întreaga ecuație de regresie pentru cazul a trei variabile, caz care se prezintă cel mai des în practica regresiei parțiale:

$$\left(x_{1\cdot 23} - M_1\right) = \frac{r_{12} - r_{13}r_{23}}{1-r^2_{23}}\cdot\frac{\sigma_1}{\sigma_2}\left(x_2 - M_2\right) + \frac{r_{13} - r_{12}r_{23}}{1-r^2_{23}}\cdot\frac{\sigma_1}{\sigma_3}\left(x_3 - M_3\right). \tag{136}$$

Ca o primă aplicație a regresiei parțiale, ne vom referi la distribuția din tabelul 109. Avem, cu notațiile întrebuințate deja:

$$M_1 = 85{,}457 \quad M_2 = 59{,}095 \quad M_3 = 167{,}026$$

$$\sigma_1 = 12{,}838 \quad \sigma_2 = 3{,}059 \quad \sigma_3 = 6{,}775$$

$$r_{12} = 0{,}859 \quad r_{23} = 0{,}381 \quad r_{13} = 0{,}377$$

și, prin urmare:

$$b_{12\cdot 3} = \frac{r_{12} - r_{13}r_{23}}{1-r^2_{23}}\cdot\frac{\sigma_1}{\sigma_2} = 3{,}517 \qquad b_{13\cdot 2} = \frac{r_{13} - r_{12}r_{23}}{1-r^2_{23}}\cdot\frac{\sigma_1}{\sigma_3} = 0{,}110.$$

Scanned with OKEN Scanner

Ecuația care ne va permite stabilirea greutății pornind de la talie și vârstă va fi, prin urmare (136):

$$x_{1\cdot23} = 85{,}457 + 3{,}517(x_2 - 59{,}095) + 0{,}110(x_3 - 167{,}026).$$

Valorile calculate pe baza acestei ecuații pentru primele zece cazuri din tabelul 109 sunt înscrise în tabelul 126. Ultimele coloane ale acestui tabel ne dau abaterile dintre valoarea statistică a greutății și valorile aceleiași variabile deduse din regresia simplă (§ 139) și regresia multiplă. Rezultatele concordă în cadrul variației statistice, deoarece nici o abatere nu întrece de două ori abaterea tip corespunzătoare:

$$\sigma_{1\cdot2} = \sigma_1\sqrt{1 - r_{12}^2} = 6{,}573$$

$$\sigma_{1\cdot23} = \sigma_1\sqrt{1 - r_{12}^2}\sqrt{1 - r_{13\cdot2}^2} = 6{,}534.$$

Prin urmare, concordanța între greutățile teoretice obținute cu ajutorul ecuației (136) și cele înregistrate efectiv este pe deplin satisfăcătoare.

TABELUL 126

Comparație între rezultatele ecuației de regresie și valorile statistice

Nr.	x_1	$x_{1\cdot23}$	$x_{1\cdot2}$	$x_1 - x_{1\cdot23}$	$x_1 - x_{1\cdot2}$
1	84,75	87,10	86,92	- 2,35	- 2,17
2	77	82,70	81,49	- 5,70	- 4,49
3	103	107,98	107,59	- 4,98	- 4,59
4	87,75	88,64	88,73	- 0,89	- 0,98
5	112,5	103,15	102,31	+ 9,35	+10,19
6	70	82,42	81,49	-12,42	- 11,49
7	91	82,04	81,49	+ 8,96	+ 9,51
8	117,25	107,32	106,83	+ 9,93	+10,42
9	82,25	81,16	79,68	+ 1,09	+ 2,57
10	95,75	88,12	87,83	+ 7,63	+ 7,92

Tabelul 109

Neînsemnata diferență dintre aceste abateri tip, precum și dintre abaterile efective din tabelul 126 vine să confirme rezultatul la care am ajuns și pe alte căi: influența vârstei asupra greutății este minimă în raport cu influența pe care o are talia asupra aceleiași variabile.

166. *Aplicație.* Trecând acum la o nouă aplicație a regresiei parțiale, vom considera tabelul 127. Avem astfel tabelele de contingență dintre: 1) greutate, 2) perimetru toracic și 3) talie, pentru un grup de 200 soldați din arma infanteriei.

Scanned with OKEN Scanner

Pentru a evita o eventuală întunecare a corelației dintre aceste variabile, din cauza suprapunerii unor colectivități eterogene, datele se referă numai la soldații de origine română proveniți din mediul rural și dintr-o regiune de șes.

Obținem pentru aceste distribuții:

$$M_1 = 67,6 \quad M_2 = 85,43 \quad M_3 = 173,73$$

$$\sigma_1 = 5,948 \quad \sigma_2 = 3,672 \quad \sigma_3 = 3,822$$

$$r_{12} = 0,601 \quad r_{23} = 0,189 \quad r_{31} = 0,340$$

de unde deducem, cu ajutorul formulelor cunoscute:

$$r_{12 \cdot 3} = 0,581 \quad r_{13 \cdot 2} = 0,289 \quad r_{32 \cdot 1} = -0,020.$$

TABELUL 127 A

Corelația dintre greutate, perimetrul toracic și talie

Perimetrul toracic în cm. (2)	Greutatea în kg (1)									
	54-57	58-61	62-65	66-69	70-73	74-77	78-81	82-85	86-89	Total
75-76	1	1	.	.	.	.	.	.	.	2
77-78	.	2	1	.	.	.	.	.	.	3
79-80	2	5	2	.	.	.	.	.	.	9
81-82	2	2	15	7	.	1	.	1	.	28
83-84	2	5	15	15	6	2	1	.	.	46
85-86	.	2	12	11	8	2	1	.	1	37
87-88	.	1	7	9	9	5	.	.	.	31
89-90	.	.	4	6	9	6	1	.	.	26
91-92	.	.	.	2	5	3	1	3	.	14
93-94	.	.	.	.	.	1	.	1	.	2
95-96	.	.	.	.	.	.	1	.	.	1
97-98	.	.	.	.	1	.	.	.	.	1
Total	7	18	56	50	38	20	5	5	1	200

Ultima dintre aceste valori pare ciudată, mai ales prin faptul că este negativă. Semnul și neînsemnata valoare a lui $r_{32 \cdot 1}$ ne îndreptățesc să afirmăm că, pentru colectivitatea considerată, nu există nici o legătură stocastică directă între talie și perimetru toracic. Într-adevăr, presupunând că valoarea teoretică a lui $r_{32 \cdot 1}$ este nulă, abaterea-tip corespunzătoare va fi $\frac{1}{\sqrt{200}} = 0,071$, așa că valoarea obținută este în limitele variației statistice. În orice caz, este greu să admitem că perimetrul toracic variind într-un sens, talia variază în sensul contrar. Această constatare a independenței dintre talie și perimetrul toracic, dacă ar fi generală, adică pentru absolut toți bărbații din satele de șes, ne-ar duce la concluzia că, pentru această categorie a populației românești,

Scanned with OKEN Scanner

dezvoltarea fizică nu se face armonios şi deci nu poate fi privită ca o bună dezvoltare. Trebuie să ţinem însă seama că datele noastre se referă la un grup de soldaţi din arma infanteriei, s-ar putea, prin urmare, ca lipsa corelaţiei directe între talie şi perimetru, constatată numeric mai sus, să se datoreze selecţionării pe arme din momentul recrutării.

Rămâne de văzut, prin studii similare şi pentru celelalte arme, dacă prima sau a doua din aceste ipoteze este cea adevărată[1].

TABELUL 127 B

Talia în cm (3)	Greutatea în kg (1)									
	54-57	58-61	62-65	66-69	70-73	74-77	78-81	82-85	86-89	Total
162-163	2	.	.	.	.	.	.	.	.	2
164-165	1	.	.	1	.	.	.	.	.	2
166-167	.	.	.	.	2	.	.	.	.	2
168-169	1	.	3	2	1	.	.	1	.	8
170-171	3	5	16	10	7	1	1	1	.	44
172-173	.	7	11	18	10	4	.	.	.	50
174-175	.	1	15	10	7	2	.	1	.	36
176-177	.	3	8	4	7	5	2	.	.	29
178-179	.	1	.	1	1	3	1	.	.	7
180-181	.	.	2	1	1	2	1	2	1	10
182-183	.	1	1	2	2	2	.	.	.	8
184-185	.	.	.	1	.	1	.	.	.	2
Total	7	18	56	50	38	20	5	5	1	200

Revenind acum la calculul ecuaţiei de regresie, vom avea:

$$b_{12} = r_{12} \frac{\sigma_1}{\sigma_2} = 0{,}974$$

şi, prin urmare, regresia greutăţii în raport numai cu perimetrul toracic va fi dată de ecuaţia:

$$x_{1\cdot 2} = 67{,}6 + 0{,}974\ (x_2 - 85{,}43). \qquad (137)$$

Regresia aceleiaşi variabile, însă în raport cu talia, este:

$$x_{1\cdot 3} = 67{,}6 + 0{,}529\ (x_3 - 173{,}73). \qquad (138)$$

Abaterile-tip ale greutăţilor evaluate prin aceste ecuaţii vor fi:

$$\sigma_{1\cdot 2} = 4{,}752 \quad \sigma_{1\cdot 3} = 5{,}591.$$

[1] *Am făcut această digresiune asupra interpretării valorii lui $r_{32\cdot 1}$ tocmai pentru a ilustra pe un exemplu viu cum lucrează metoda statistică. Faptele de mai sus ar fi fost greu puse în lumină fără ajutorul acesteia din urmă.*

Scanned with OKEN Scanner

Și aceste valori ca și coeficienții r_{12}, r_{13} ne arată că, pentru clasa examinată, obținem evaluări mult mai bune pentru greutate pornind de la perimetrul toracic, decât dacă am considera talia ca element determinant al primei valori.

TABELUL 127 C

Talia în cm (3)	Perimetrul toracic în cm (2)												
	75-76	77-78	79-80	81-82	83-84	85-86	87-88	89-90	91-92	93-94	95-96	97-98	Total
162-63	1	.	.	.	1	.	.	.	.	.	.	.	2
164-65	.	.	1	.	.	.	1	.	.	.	.	.	2
166-67	.	.	.	.	.	.	.	2	.	.	.	.	2
168-69	.	.	.	2	2	.	2	1	1	.	.	.	8
170-71	.	.	3	9	18	7	3	2	.	1	1	.	44
172-73	1	.	2	7	7	14	8	6	5	.	.	.	50
174-75	.	2	1	7	7	3	4	7	3	1	.	1	36
176-77	.	1	2	3	4	7	7	4	1	.	.	.	29
178-79	.	.	.	.	1	.	5	1	.	.	.	.	7
180-81	.	.	.	.	2	3	.	2	3	.	.	.	10
182-83	.	.	.	.	4	2	1	.	1	.	.	.	8
184-85	.	.	.	.	.	1	.	1	.	.	.	.	2
Total	2	3	9	28	46	37	31	26	14	2	1	1	200

TABELUL 128

Comparație între rezultatele ecuației de regresie și datele statistice

Talia în cm	Perimetrul toracic în cm						
	75-78	79-82	83-86	87-90	91-94	95-98	$x_{1\cdot3}$
162-165	54 **55,81** (4,38)	56 **59,41** (4,38)	54 **63,02** (4,38)	68 **66,63** (4,38)	.	.	58 **62,19** (2,80)
166-169	.	68,5 **60,94** (3,09)	66 **64,55** (3,09)	67,6 **68,16** (1,96)	70 **71,77** (4,38)	.	67,7 **64,30** (1,77)
170-173	59 **58,73** (4,38)	63,57 **62,33** (0,95)	65,89 **65,94** (0,65)	68,26 **69,55** (1,00)	72,33 **73,16** (1,79)	80 **76,77** (4,38)	66,34 **66,42** (0,58)
174-177	61 **60,20** (2,53)	62,77 **63,80** (1,21)	67,48 **67,41** (0,95)	69,91 **71,02** (0,93)	75 **74,63** (1,96)	72 **78,24** (4,38)	67,71 **68,54** (0,69)
178-181	.	.	70,83 **68,87** (1,79)	72,5 **72,48** (1,55)	82,67 **76,09** (2,53)	.	73,71 **70,65** (1,36)
182-185	.	.	67,71 **70,34** (1,65)	71 **73,95** (3,09)	74 **77,56** (4,38)	.	69 **72,77** (1,77)
$x_{1\cdot2}$	59,20 **58,90** (2,13)	63,42 **62,80** (0,78)	66,66 **66,69** (0,52)	69,53 **70,59** (0,63)	75,06 **74,49** (1,19)	76 **78,38** (3,36)	

Tabelul 127

Regresia multiplă a greutății în raport cu perimetrul toracic și cu talia va fi:

$$x_{1\cdot23} = 67{,}6 + 0{,}902\,(x_2 - 85{,}43) + 0{,}366\,(x_3 - 173{,}73). \qquad (139)$$

Abaterea lui $x_{1\cdot23}$ va fi:

Scanned with OKEN Scanner

$$\sigma_{1\cdot23} = \sigma_{1\cdot2}\sqrt{1 - r^2_{13\cdot2}} = 4{,}752 \times 0{,}957 = 4{,}375$$

mult mai mică decât $\sigma_{1\cdot2}$ şi $\sigma_{1\cdot3}$.

Pentru o comparaţie între valorile statistice ale lui x_1 şi cele ale lui $x_{1\cdot23}$ am construit tabelul 128.

Pentru a evita ca o cază să cuprindă un număr prea restrâns de cazuri, fapt care ar putea pune într-o poziţie exagerată valorile excepţionale, clasificarea din acest tabel s-a făcut în raport cu intervale egale cu 4 cm, atât pentru talie, cât şi pentru perimetrul toracic, ceea ce s-a obţinut prin gruparea la un loc a câte două linii şi două coloane din ultimul tabel (127 C).

Prima cifră înscrisă în oricare din cazele acestui tabel reprezintă media aritmetică a greutăţii statistice a elementelor corespunzătoare acestei caze, cifra imediat următoare (în caractere groase) este greutatea teoretică.

Astfel, pentru caza subliniată, avem 22 de soldaţi care au perimetrul toracic cuprins între 87-90 cm şi talia între 174-177 cm. Greutăţile acestor 22 de soldaţi sunt în kg[1]:

66, 66, 73, 62, 73, 67, 65, 73, 74, 73, 73, 73, 68, 70, 76, 65, 68, 66, 67, 76, 70, 74

de unde obţinem greutatea medie statistică 69,91. Pentru aceeaşi cază, valorile medii ale perimetrului şi taliei sunt respectiv:

$$x_2 = 88{,}5\ cm \qquad x_3 = 175{,}5$$

adică mijloacele intervalelor corespunzătoare. Din ecuaţia (139) deducem pentru greutate valoarea medie teoretică:

$$x_{1\cdot23} = 67{,}6 + 0{,}902 \times 3{,}07 + 0{,}366 \times 1{,}77 = 71{,}02.$$

La fel am obţinut mediile statistice şi teoretice pentru regresiile simple, pornind de la ecuaţiile (137) şi 138). Rezultatele sunt înscrise în coloanele marginale.

Numerele din paranteze reprezintă abaterea-tip a diferenţei dintre valoarea teoretică şi valoarea statistică. Dacă fiecare cază nu ar cuprinde decât un singur element, abaterea-tip corespunzătoare ar fi egală cu $\sigma_{1\cdot23} = 4{,}375$; cum avem însă caze cu un număr mai mare de elemente, vom compara de fapt nişte medii între ele şi nu nişte valori individuale. Dacă $\sigma_{1\cdot23}$ este abaterea tip a valorilor individuale, abaterea-tip a mediei deduse din n astfel de valori va fi (§ 102): $\frac{\sigma_{1\cdot23}}{\sqrt{n}}$. Pentru caza considerată mai sus, avem:

[1] *Pentru a nu îngreuna prea mult această carte cu un tabel în plus, care nu ni se pare de o prea mare necesitate, nu mai publicăm seriile individuale ale tuturor valorilor care au dat naştere la distribuţiile din tabelul 127.*

Scanned with OKEN Scanner

$$\frac{4,375}{\sqrt{22}} = 0,93.$$

Abaterile-tip ale regresiilor simple sunt calculate după aceeași metodă, pornind însă de la $\sigma_{1\cdot 2}$ și $\sigma_{1\cdot 3}$.

Din comparația diferențelor efective dintre x_1 și $x_{1\cdot 23}$ cu abaterile-tip respective, se constată că numai în două cazuri din cele 26 de caze pentru care am putut face comparația diferențele statistice întrec de 2 ori această abatere. Referindu-ne la cele spuse în § 85, ar fi trebuit ca în mod teoretic să avem $2 \times 0,0456 = 1,19$ cazuri în care să avem o abatere mai mare decât de două ori abaterea tip. Prin urmare, concordanța între greutățile teoretice obținute cu ajutorul ecuației (139) și cele înregistrate efectiv este pe deplin satisfăcătoare.

167. *Corelația rangurilor.* Înainte de a încheia acest capitol, ne vom mai opri asupra unei alte metode de măsură a corelației. Este vorba despre *corelația rangurilor*. Ideea care stă la baza acestei corelații a fost expusă pentru prima oară de dr. C. SPEARMAN[1], însă construcția logică a coeficientului de corelație a rangurilor, împreună cu rectificarea unor afirmații eronate ale celui dintâi, aparține lui K. PEARSON[2].

Considerând seriile de valori a două variabile, putem, aranjând aceste valori într-o ordine descrescătoare sau crescătoare, atribui fiecăreia un rang. Dacă fenomenele considerate sunt în legătură stocastică directă, atunci cazurile care s-au clasificat printre cele dintâi în raport cu prima variabilă vor fi printre primele și în clasificarea în raport cu cea de-a doua variabilă și viceversa. Dacă legăturii stocastice îi corespunde o corelație negativă, rangurile vor fi răsturnate de la o clasificare la alta.

Astfel, dacă va exista o corelație între aptitudinile școlarilor pentru matematici și cele pentru științele fizico-chimice, aceia dintre școlari care se vor clasifica printre primii la o probă de matematici vor ocupa primele ranguri și după o probă de fizico-chimie. Prin urmare, o comparație a rangurilor ne va da posibilitatea să punem în evidență gradul legăturii stocastice dintre două fenomene.

Să ne referim la tabelul 129 în care am clasificat județele Vechiului Regat, în afară de Dobrogea, în raport cu natalitatea și mortalitatea medie pentru anii 1901-1903.

Făcând ordonarea cazurilor examinate după mărimea lor, observăm că de la început se ridică o ambiguitate: două sau mai multe cazuri se pot prezenta cu aceeași valoare, fapt care nu ne mai permite să atribuim fiecărui element un rang unic și bine determinat. Dificultatea aceasta se poate soluționa în două chipuri:

a) *Prin metoda rangului mijlociu.* Dacă două sau mai multe elemente au aceeași valoare, se va atribui fiecăruia un rang egal cu media rangurilor pe care le ocupă acestea. Astfel, valorilor:

---

[1] *Dr. C. Spearman, **A Foot-rule for Measuring Correlation**, British Journal of Psychology, vol. 2, p. 89.*

[2] *K. Pearson, **On Further Methods of Determining Correlation**, Drapers' Company Research Memoirs, Biometric Series IV.*

Scanned with OKEN Scanner

2, 4, 4, 4, 5, 10, 10

li se vor atribui, respectiv, rangurile:

1; 3; 3; 3; 5; 6,5; 6,5.

b) *Prin metoda rangului prim*. Prin această metodă, se atribuie fiecăruia din elementele egale rangul pe care îl ocupă primul din acestea. În cazul de mai sus, rangurile vor fi, prin urmare:

1, 2, 2, 2, 5, 6, 6.

Observăm că atunci când trecem de la o grupă de elemente egale la următoarea începem, în oricare din metode, să socotim rangul primului element ce urmează ca și cum toate elementele precedente ar fi fost neegale.

TABELUL 129

Calculul coeficientului de corelație al rangurilor pentru natalitatea și mortalitatea pe județe, în România (1901 - 1903)

Județul	Natalitate	Mortalitate	Metoda rangului mediu			Metoda rangului prim		
			v_1	v_2	$(v_1 - v_2)^2$	v_1	v_2	$(v_1 - v_2)^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Argeș	35,0	22,1	1	1	-	1	1	-
Bacău	38,4	26,8	11	13,5	6,25	11	13	4
Botoșani	39,1	27,6	13	20	49	13	19	36
Brăila	43,7	27,4	27	18	81	27	18	81
Buzău	41,4	27,2	22,5	15,5	49	22	15	49
Covurlui	39,6	29,7	16	30	196	16	30	196
Dâmbovița	41,4	25,4	22,5	9	182,25	22	9	169
Dolj	39,2	28,0	14,5	24	90,25	14	23	81
Dorohoi	40,2	26,8	19	13,5	30,25	19	13	36
Fălciu	42,8	27,2	26	15,5	110,25	26	15	121
Gorj	40,1	24,9	18	8	100	18	8	100
Ialomița	46,7	27,6	30	20	100	30	19	121
Iași	37,8	28,7	10	27	289	10	27	289
Ilfov	37,2	26,6	5,5	12	42,25	5	12	49
Mehedinți	37,0	24,3	4	4	-	4	4	-
Muscel	36,8	22,9	3	2	1	3	2	1
Neamț	36,6	24,7	2	6	16	2	6	16
Olt	37,7	24,0	8,5	3	30,25	8	3	25
Prahova	39,7	24,5	17	5	144	17	5	144
Putna	37,3	28,8	7	28	441	7	28	441
R. Sărat	40,5	27,3	20	17	9	20	17	9
Roman	39,2	27,6	14,5	20	30,25	14	19	25

Scanned with OKEN Scanner

(continuare tab. 129)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Romanați	41,7	29,4	24,5	29	20,25	24	29	25
Suceava	37,7	25,9	8,5	10	2,25	8	10	4
Tecuci	41,0	28,0	21	24	9	21	23	4
Teleorman	45,1	28,6	28,5	26	6,25	28	26	4
Tutova	39,0	27,9	12	22	100	12	22	100
Vaslui	41,7	26,2	24,5	11	182,25	24	11	169
Vâlcea	37,2	24,8	5,5	7	2,25	5	7	4
Vlașca	45,1	28,0	28,5	24	20,25	28	23	25
					2339,50			2328

PEARSON recomandă cea de-a doua metodă. Se pare, totuși, că cea dintâi este cea mai logică în cazul variabilelor continue, cea de a doua, în cazul variabilelor discontinue. Într-adevăr, este evident că într-o clasificare a unui grup de flori în raport cu numărul petalelor, exemplarele care prezintă 5 petale trebuie să ocupe rangul pe care îl ocupă primul dintre ele, deoarece nimic nu le deosebește unul de altul. Când însă variabila este continuă, egalitatea dintre un grup de valori este, cu mici excepții, numai aparentă. Calcularea uneia sau a două cifre zecimale în plus va face ca egalitatea dintâi să dispară, iar elementele se vor ordona atunci fără nici o ambiguitate. Ignorând însă valoarea zecimalelor de care vorbim, vom atribui tuturor elementelor același rang și, pentru a comite o eroare cât mai mică, îl vom lua egal cu rangul mijlociu.

O dată rangurile determinate prin una sau alta din metode, se calculează pentru fiecare caz în parte pătratul $(\nu_1 - \nu_2)^2$, unde ν_1, ν_2 sunt valorile rangului aceluiași element în cele două clasificări.

Coeficientul de corelație a rangurilor sub forma propusă de K. PEARSON este în urmă calculat pornind de la formula:

$$\rho = 1 - \frac{6\sum(\nu_1 - \nu_2)^2}{N(N^2 - 1)} \qquad (140)$$

coeficient care variază între -1 și +1, atingând aceste valori în cazul unei corelații perfecte a rangurilor. Adică ρ = +1, când fiecare element ocupă același rang în ambele serii, și ρ = -1, când elementul de rangul i dintr-o serie ocupă rangul $N - i + 1$ în cealaltă (rangurile vor fi deci complet răsturnate de la o serie la alta).

În cazul seriilor din tabelul 129, avem :

Prin metoda rangului mijlociu: $\rho = 1 - \frac{6 \times 2339,5}{30 \times 899} = 0,480$

Prin metoda rangului prim: $\rho = 1 - \frac{6 \times 2328}{30 \times 899} = 0,482$.

(continuare tab. 129)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Romanați	41,7	29,4	24,5	29	20,25	24	29	25
Suceava	37,7	25,9	8,5	10	2,25	8	10	4
Tecuci	41,0	28,0	21	24	9	21	23	4
Teleorman	45,1	28,6	28,5	26	6,25	28	26	4
Tutova	39,0	27,9	12	22	100	12	22	100
Vaslui	41,7	26,2	24,5	11	182,25	24	11	169
Vâlcea	37,2	24,8	5,5	7	2,25	5	7	4
Vlașca	45,1	28,0	28,5	24	20,25	28	23	25
					2339,50			2328

PEARSON recomandă cea de-a doua metodă. Se pare, totuși, că cea dintâi este cea mai logică în cazul variabilelor continue, cea de a doua, în cazul variabilelor discontinue. Într-adevăr, este evident că într-o clasificare a unui grup de flori în raport cu numărul petalelor, exemplarele care prezintă 5 petale trebuie să ocupe rangul pe care îl ocupă primul dintre ele, deoarece nimic nu le deosebește unul de altul. Când însă variabila este continuă, egalitatea dintre un grup de valori este, cu mici excepții, numai aparentă. Calcularea uneia sau a două cifre zecimale în plus va face ca egalitatea dintâi să dispară, iar elementele se vor ordona atunci fără nici o ambiguitate. Ignorând însă valoarea zecimalelor de care vorbim, vom atribui tuturor elementelor același rang și, pentru a comite o eroare cât mai mică, îl vom lua egal cu rangul mijlociu.

O dată rangurile determinate prin una sau alta din metode, se calculează pentru fiecare caz în parte pătratul $(\nu_1 - \nu_2)^2$, unde ν_1, ν_2 sunt valorile rangului aceluiași element în cele două clasificări.

Coeficientul de corelație a rangurilor sub forma propusă de K. PEARSON este în urmă calculat pornind de la formula:

$$\rho = 1 - \frac{6\sum(\nu_1 - \nu_2)^2}{N(N^2 - 1)} \qquad (140)$$

coeficient care variază între -1 și +1, atingând aceste valori în cazul unei corelații perfecte a rangurilor. Adică ρ = +1, când fiecare element ocupă același rang în ambele serii, și ρ = -1, când elementul de rangul i dintr-o serie ocupă rangul $N - i + 1$ în cealaltă (rangurile vor fi deci complet răsturnate de la o serie la alta).

În cazul seriilor din tabelul 129, avem :

Prin metoda rangului mijlociu: $\rho = 1 - \frac{6 \times 2339,5}{30 \times 899} = 0,480$

Prin metoda rangului prim: $\rho = 1 - \frac{6 \times 2328}{30 \times 899} = 0,482$.

Aceste două valori nu diferă simţitor una de alta. Lucrurile nu se vor petrece la fel în cazul unei serii conţinând un număr mai restrâns de valori. Atunci, cele două metode ne vor conduce la rezultate diferite şi va fi nevoie să ţinem seamă de consideraţiile expuse mai sus pentru a stabili care anume din rezultate îl vom considera valabil.

În cazul când variabilele considerate se distribuie normal (ceea ce se întâmplă pentru coeficienţii de natalitate şi mortalitate), valoarea lui ρ ne permite să determinăm coeficientul de corelaţie pearsonian prin relaţia:

$$r = 2\sin\left(\frac{\pi}{6}\rho\right). \qquad (141)$$

Pentru tabelul 129, avem:

$$\frac{\pi}{6} \times 0{,}48 \approx 14°24'$$

de unde, cu ajutorul unei table de valori naturale ale liniilor trigonometrice, obţinem:

$$\sin 14°24' = 0{,}2487$$

deci:

$$r = 0{,}497.$$

Valoarea aceasta a coeficientului r nu are însă aceeaşi precizie ca cea obţinută direct, abaterea ei tip:

$$\frac{1{,}047(1-r^2)}{\sqrt{N}}$$

fiind mai mare decât aceea a coeficientului r (§ 144). Valoarea acestuia din urmă, calculată prin metoda expusă în Cap. XI, este:

$$r = 0{,}500$$

valoare care, nediferind simţitor de cea dedusă din (141), confirmă normalitatea distribuţiei coeficienţilor de natalitate şi mortalitate[1].

---

[1] *Această corelaţie netăgăduită dintre natalitate şi mortalitate ar putea părea ciudată la prima vedere, fenomenele nefiind plauzibil legate într-un mod cauzal oarecare. Faptul se explică însă. România fiind o ţară cu o mortalitate infantilă foarte ridicată - aproximativ 30% din totalul deceselor îl furnizează copiii sub 1 an - urmează că acolo unde sunt naşteri mai numeroase, vor fi şi decese mai multe. Sunt dator pentru această interpretare dlui dr. S. MANUILĂ, directorul Institutului de Demografie şi Recensământ.*

Scanned with OKEN Scanner

PARTEA a IV-a

Scanned with OKEN Scanner

Capitolul XIII

Analiza seriilor în timp

168. *Generalităţi.* Cititorul a observat fără îndoială că, în afară de mici excepţii, seriile în timp nu au fost privite în cadrul capitolelor precedente nici sub unghiul unei analize descriptive şi nici din punctul de vedere al procedeelor de prelucrare tehnică. Acest fapt nu este o simplă întâmplare şi nici nu trebuie considerat ca o consecinţă a unei anumite ordini de prezentare a materiei. Tratarea separată a seriilor în timp este impusă de însăşi structura acestora.

Toată teoria distribuţiilor de frecvenţe are la bază un postulat - postulatul stabilităţii frecvenţelor. Pe acest postulat sau mai degrabă pe această stabilitate se clădeşte întreaga teorie a distribuţiilor de frecvenţe, care îmbracă astfel o formă ştiinţifică netăgăduită.

O analiză a seriilor în timp ne va arăta imediat că, din cauza felului cum acestea se prezintă, nu mai putem desprinde nici un element de permanenţă în legătură cu structura lor. Suntem nevoiţi, în acest caz, să ne conducem, pentru studiul seriilor în timp, de aspectul formal al acestora. Orice desconsiderare a diferenţei fundamentale dintre structura distribuţiilor de frecvenţă şi aceea a seriilor în timp va constitui originea principală a oricăruia din aşa-numitele eşecuri ale metodei statistice.

O serie statistică în timp oglindeşte printr-un şir de date *evoluţia* unui fenomen colectiv. În majoritatea cazurilor, această evoluţie nu are nimic permanent în felul cum se desfăşoară şi orice încercare de a o descrie printr-o formulă oarecare - matematică sau nu - va fi implicit zadarnică. Această constatare îmbrăţişează o chestiune mult mai generală decât simpla analiză a seriilor în timp: ea priveşte problema determinismului sub forma lui cea mai netă. Nu vom intra în discuţia problemei astfel pusă, ci ne vom retrage pe poziţia statistică a formulării evoluţiei fenomenelor.

Este drept că pentru anumite categorii de colectivităţi *organizate* evoluţia lor are un sens precis care se desprinde din forma statistic asemănătoare de la un caz la altul. Unul dintre cele mai clasice exemple îl constituie evoluţia dezvoltării unui organism viu. Nu este nevoie să definim ce înţelegem prin creşterea în înălţime sau în greutate a unui copil de la naştere şi până la maturitate. Vom vorbi fără nici o reticenţă despre evoluţia unei plantaţii şi chiar a unei populaţii oarecare. Proba permanenţelor acestor forme de evoluţii o constituie înseşi planurile de exploatare pentru păduri, tablele de mortalitate pentru populaţii, dar mai ales faptul că pe acestea din urmă vine să se

Scanned with OKEN Scanner

investească cu preferinţă elementul cel mai capricios în alegerea încrederii sale: capitalul. Iar dacă există faliţi şi printre exploatatorii de păduri, nu faptul că pădurea nu a putut furniza la un moment oarecare cantitatea de lemn prevăzută este cauza.

TABELUL 130

Consumul de cărbuni în Franţa (1876 - 1910)

Anii	Mii tone	Anii	Mii tone	Anii	Mii tone	Anii	Mii tone	Anii	Mii tone
1876	24.472	1883	32.439	1890	36.653	1897	41.841	1904	47.030
1877	24.144	1884	30.941	1891	36.573	1898	43.295	1905	48.669
1878	24.555	1885	30.035	1892	36.516	1899	42.228	1906	51.782
1879	25.332	1886	29.619	1893	36.379	1900	48.803	1907	55.104
1880	28.846	1887	31.191	1894	38.000	1901	46.773	1908	54.740
1881	29.445	1888	32.674	1895	38.640	1902	44.810	1909	56.350
1882	31.025	1889	33.511	1896	39.995	1903	48.180	1910	56.536

Extras din ***Annuaire statistique de la France****, 1932, p. 145.*

Se întâmplă însă ca interesul general să fie atras mai ales către evoluţia fenomenelor economico-financiare. Aceste fenomene fac parte din categoria acelora pe care le voi numi *organizabile*, adică asupra cărora omul poate să intervină prin măsuri de ordin modificator, când într-un sens, când într-altul. Cum se mai poate vorbi de evoluţia unor astfel de fenomene ca despre o noţiune bine circumscrisă? Cine mai poate avea sentimentul de încredere ştiinţifică în permanenţa unor legi de evoluţie formulate în urma analizei pe o perioadă limitată de timp a unui astfel de fenomen? Nu este logic să ne aşteptăm ca formula găsită azi să fie infirmată în viitor din cauza intervenţiei omului?

Mai sunt, de asemenea, cazuri de fenomene colective organizate, dar şi organizabile în acelaşi timp. Mortalitatea de tuberculoză face parte din această categorie. Recunoaştem în evoluţia acestui fenomen suficiente elemente permanente, însă o politică sanitară diferită de cea a primei perioade de observaţie poate schimba complet alura historiogramei respective în perioada imediat următoare. De fapt, se poate întâmpla ca fenomenul să dispară chiar dacă, de pildă, s-ar descoperi un vaccin contra tuberculozei.

Iată câteva aspecte ale problemei analizei seriilor în timp, care ne indică încă de pe acum care va fi calea pe care o vom adopta pentru studiul unei aşa de spinoase chestiuni.

169. *Componentele mişcării în timp.* Să trecem acum la un examen mai amănunţit al aspectelor evoluţiei unui fenomen colectiv. Să considerăm, în primul rând, seria din tabelul 52. Historiograma corespunzătoare (graficul 58) ne arată existenţa unei evoluţii nete a unei populaţii limitate la un teritoriu aproape constant

Scanned with OKEN Scanner

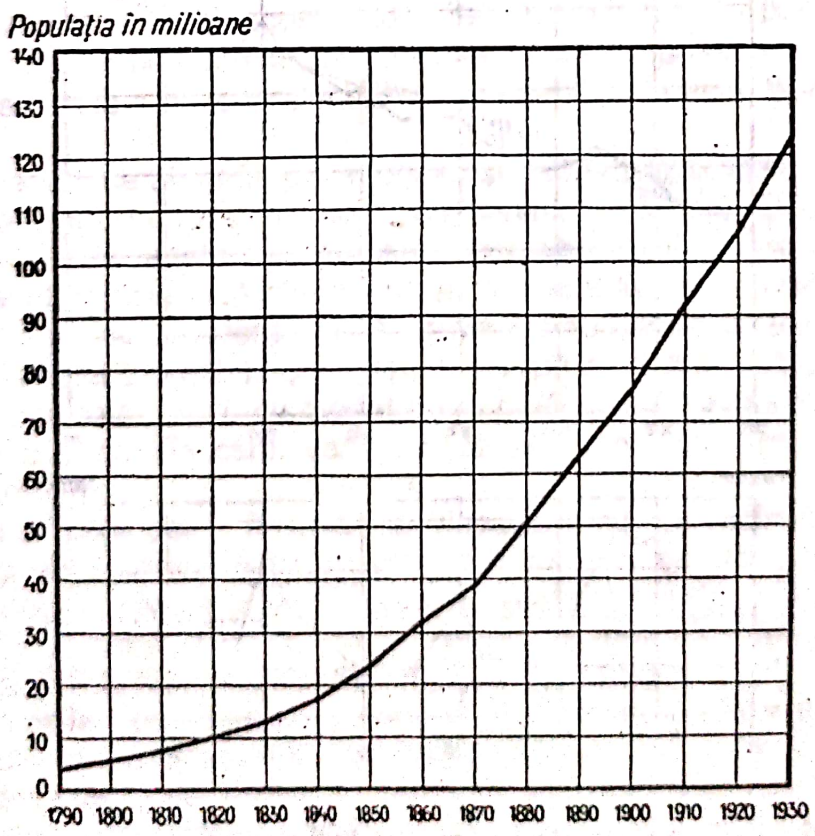
de-a lungul timpului. Permanența formei acestei evoluții se poate constata prin compararea graficelor respective pentru creșterea populațiilor din diferite țări.

O tendință evolutivă se poate desprinde și din examenul graficului 59 A (tabelul 130), cu toate că în acest din urmă caz degajarea acesteia din aspectul general al historiogramei se face anevoios.

Printr-o abstractizare ale cărei baze logice sunt destul de șubrede, putem privi linia dreaptă trasată pe historiogramă ca reprezentând tendința generală a evoluției fenomenului considerat. Evident că această tendință ar putea fi *descrisă* tot așa de bine și printr-o altă linie curbă sau chiar frântă. Nu există mijloace care să ne permită o alegere în această direcție. Și acest lucru se explică prin lipsa unei definiții complete a ceea ce noi înțelegem prin tendința evolutivă a unui fenomen.

Această tendință a fost totuși denumită cu termenul de *tendință seculară sau de lungă durată*. Nici unul din acești termeni nu pare însă nimerit ales, deoarece, departe de a oglindi relativitatea noțiunii ce o exprimă, îi atribuie o permanență - de lungă durată sau seculară - care, după cum am văzut, este inexistentă în majoritatea cazurilor.

Cu aceste rezerve, noțiunea de tendință de lungă durată este totuși utilă procedeelor de investigație și nu poate lipsi din nici o expunere completă a metodelor statisticii matematice.

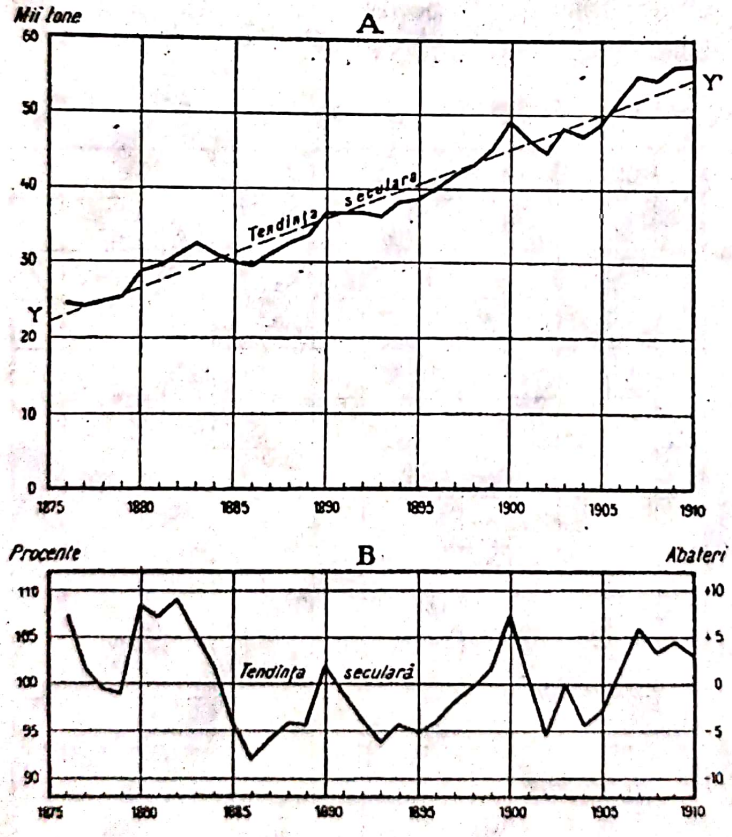


Tabelul 52

Graficul 58. Evoluția populației din Statele Unite (1790 - 1930).

Scanned with OKEN Scanner

Trecând acum la oscilațiile liniei reprezentând seria efectivă în jurul liniei de lungă durată, nu ne putem împiedica de a observa că și acestea cuprind o permanență formală. Aceste oscilații nu se succed la întâmplare, ci trecerea dintr-o parte în alta a liniei de lungă durată se face la anumite intervale, în interiorul cărora linia efectivă rămâne aproape mereu de aceeași parte a celei dintâi. Acest fapt dă mișcării liniei efective alura unor ondulații, din care cauză i s-a și dat acestei mișcări numele de *mișcare ondulatorie*. Termenul consacrat însă este acela de *mișcare ciclică*. Acesta din urmă prezintă totuși marele neajuns de a atribui mișcării ondulatorii permanența recurentă și regulată a ciclurilor. Într-adevăr, prin însăși definiția lor, ciclurile se repetă în mod identic la același interval și cu aceeași amplitudine, pe când mișcarea ondulatorie este departe de a fi chiar așa de regulată. Introducerea și menținerea termenului de variație ciclică se datorează primelor cercetări în această direcție. JUGLAR și STANLEY JEVONS în secolul trecut, prof. H. L. MOORE la începutul secolului XX au încercat să evidențieze o regularitate ciclică a ondulațiilor fenomenelor economice. Timpul și cercetările mai moderne au arătat că acestea sunt departe de a fi reprezentate prin cicluri; termenul a rămas însă. Ar fi poate o impietate să-l înlăturăm din vocabularul științific.



Tabelul 130

Graficul 59. Tendința de lungă durată și mișcarea ondulatorie a consumului de cărbuni în Franța (1876 - 1910).

TABELUL 131

Consumul lunar de ciment în România (1927 - 1932)

Lunile	TONE					
	1927	1928	1929	1930	1931	1932
Ianuarie	2.140	2.590	1.810	4.590	3.070	1.723
Februarie	6.920	6.360	1.980	9.230	5.130	1.481
Martie	19.520	19.480	9.310	28.610	11.890	3.920
Aprilie	31.210	22.620	26.840	23.310	15.620	19.210
Mai	32.970	35.300	36.880	35.210	29.620	23.820
Iunie	33.600	40.490	41.720	32.930	26.920	27.710
Iulie	31.110	37.870	41.000	33.430	29.100	31.410
August	39.970	40.420	42.420	36.030	29.280	33.480
Septembrie	36.300	41.576	40.050	30.630	25.890	34.380
Octombrie	33.510	35.870	31.600	25.180	20.320	25.830
Noiembrie	15.640	20.140	22.840	14.720	10.330	6.060
Decembrie	3.180	4.590	11.130	7.430	1.790	8.367

Date extrase din ***Buletinul Institutului de Conjunctură****, anul I, nr. 1-2, p. 56.*

Mişcarea ondulatorie poate fi de multe ori perfect ciclică - cum este aceea a petelor solare - însă această identitate nu se produce pentru nici unul din fenomenele economico-financiare analizate până în ziua de azi.

Mişcarea ciclică a consumului de cărbuni din Franţa este reprezentată de graficul 59 B.

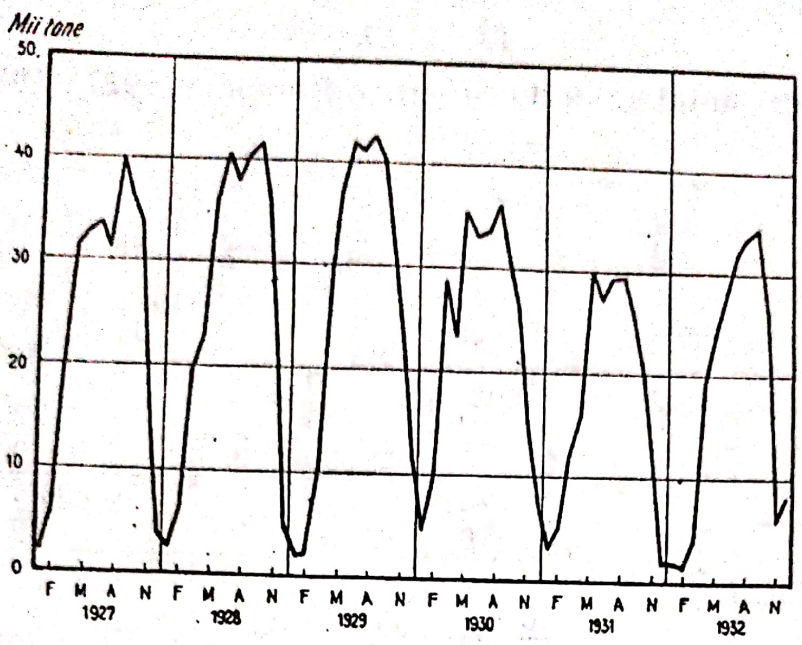
În fine, variaţii de alt ordin pot interveni şi modifica evoluţia unui fenomen. Dintre cele mai importante, vom menţiona variaţia sezonală. Diferenţa dintre anotimpuri, obiceiurile legate de anumite epoci ale anului vin să imprime mai tuturor fenomenelor sociale un ritm diferit de la o lună la alta. Existenţa variaţiei sezonale este mult mai evidentă decât a celor două mişcări menţionate mai sus. Seria din tabelul 5 prezintă o variaţie sezonală netă datorată obiceiurilor şi regulilor religioase în legătură cu celebrarea căsătoriilor, în timp ce tabelul 131 ne înfăţişează o serie ale cărei variaţii sezonale sunt determinate de cauze specifice anotimpurilor.

Această analiză poate fi dusă mai departe; vom da astfel peste variaţii săptămânale şi chiar zilnice (§12). Acestea din urmă nu prezintă însă decât un interes prea special, aşa că le vom lăsa de o parte. De altfel, chiar publicaţiile statistice cele mai amănunţite nu conţin decât cel mult date lunare, variaţiile săptămânale şi zilnice fiind prin aceasta imposibil de studiat.

Consideraţiile precedente ne conduc la descompunerea evoluţiei unui fenomen colectiv în următoarele componente:

a) *Tendinţa de lungă durată*;
b) *Mişcarea ondulatorie (ciclică)*;
c) *Variaţia sezonală*.

Scanned with OKEN Scanner



Tabelul 131

Graficul 60. Consumul lunar de ciment din România.

Problema descompunerii evoluției unui fenomen în componentele menționate prezintă un dublu interes:

1. *Practic.* Sunt o seamă de măsuri pe care statul sau chiar particularii trebuie să le ia din vreme pentru a adapta mersul întreprinderilor lor ritmului variației sezonale în scopul de a obține maximum de beneficiu. În ceea ce privește tendința de lungă durată, dar mai ales mișcarea ondulatorie, deși poate mai interesante de folosit în practică, nu pot face încă obiectul unei încrederi generale.

2. *Teoretic.* Dacă dorim însă ca, printr-o analiză a datelor statistice, să stabilim numai anumite legături între fenomenele economice, atunci este neapărată nevoie ca evoluția fenomenelor considerate să fie descompusă în cele trei componente fundamentale. Legile care pot fi adevărate pentru tendințele de lungă durată pot să fie contrazise de mișcarea ciclică a fenomenelor. Pentru a nu menționa decât un exemplu, ne vom referi la legătura dintre mișcarea prețurilor și aceea a salariilor reale[1]: variația acestora se face în același sens dacă se consideră mișcarea de lungă durată, în sens contrar. pentru variațiile ciclice.

De multe ori o legătură netăgăduită dintre două fenomene nu poate fi pusă statistic în evidență decât după descompunerea evoluțiilor lor.

170. *Determinarea grafică a tendinței de lungă durată.* Mișcarea de lungă durată se poate determina prin trei procedee diferite.

---

[1] *Salariul real este raportul dintre salariul efectiv și indicele costului vieții.*

Cel mai simplu dintre acestea constă în trasarea cu mâna liberă a unei linii *line*[1] care să urmărească de aproape alura generală a historiogramei. Linia aceasta este cunoscută în literatura saxonă sub numele de *french curve*, din cauza preferinţei pe care a avut-o şi pe care încă o are şcoala franceză pentru acest mod de determinare a tendinţei de lungă durată.

Trebuie recunoscut de la început că acest procedeu prezintă dezavantajul, care nu poate fi ignorat, de a introduce în operaţia de determinare a mişcării seculare nu numai diferenţe de la o persoană la alta, dar chiar pentru o aceeaşi persoană, diferenţe de la o operaţie la alta. Nu este nevoie să mai accentuăm aceste lucruri evidente.

Neajunsul cel mare pe care îl prezintă însă această metodă constă mai ales din faptul că, oricât de simplă ar părea la prima vedere, aplicarea ei cere nu numai o îndemânare oarecare, ci şi o intuiţie a tendinţei de lungă durată, care, din nenorocire, nu poate fi însuşită decât în urma unui continuu şi repetat contact cu astfel de probleme. Iată deci cum, de la început, metoda nu poate fi accesibilă începătorilor.

Cu toate acestea, nu este mai puţin adevărat că aceasta este singura metodă care corespunde întocmai concepţiei pe care o avem despre tendinţa de lungă durată. Neajunsurile metodei ţin de imprecizia cuprinsului noţiunii la care se referă.

171. *Metoda mediilor mobile.* O a doua metodă de determinare a tendinţei de lungă durată este cunoscută sub numele de metoda mediilor mobile. Principiul pe care se bazează această metodă este acelaşi cu cel întâlnit mai sus (§ 55) şi pe care l-am putea numi *principiul compensaţiei*. Dacă există abateri de natură ciclică de la tendinţa de lungă durată, făcând media mai multor termeni consecutivi, vom avea toate şansele ca aceste abateri să se compenseze între ele, astfel încât rezultatul final să nu mai cuprindă decât elementul căutat: tendinţa de lungă durată.

Procedeul de calcul este exemplificat prin tabelul 132.

Fiecare din termenii seriei este înlocuit prin media aritmetică a unui număr impar de termeni pentru care termenul considerat ocupă poziţia centrală. Astfel, valoarea tendinţei de lungă durată corespunzătoare anului 1883 va fi, în cazul unei medii mobile purtând asupra unui număr de cinci ani, media aritmetică a datelor referitoare la anii 1881, 1882, 1883, 1884, 1885.

Numărul de ani adoptat pentru aplicarea procedeului mediilor mobile este arbitrar şi poate fi foarte bine unul oarecare din numerele impare 3, 5, 7, 9 etc. Regularitatea ciclică pe care JUGLAR şi JEVONS credeau că au descoperit-o pentru fenomenele economice şi pe care o apreciau sub forma unui ciclu de 9 ani a făcut să se adopte de preferinţă media mobilă pe 9 ani. Este clar că, în cazul unor cicluri de nouă ani, fiecare grupă de nouă termeni consecutivi va cuprinde toate fazele unui atare ciclu şi, prin urmare, compensaţia de care vorbeam mai sus va deveni o certitudine.

---

[1] Smooth, *în engleză. O linie lină este întotdeauna continuă, reciproca nu este însă adevărată şi, prin urmare, aceste două aspecte nu trebuie confundate unul cu altul.*

Scanned with OKEN Scanner

Metoda mediilor mobile prezintă în acelaşi timp şi avantaje, şi neajunsuri.

Primul neajuns este de ordin ideologic: prin însăşi aplicarea acestui procedeu pentru determinarea tendinţei de lungă durată înseamnă că adoptăm implicit pentru această noţiune o definiţie diferită de cea pe care o cunoaştem deja. A defini mişcarea de lungă durată ca mişcarea pe care ar avea-o fenomenul în cazul când nu ar exista variaţia ciclică şi a accepta că ea este indicată de mediile aritmetice ale unor termeni consecutivi din seria dată sunt două lucruri care nu pot fi identificate în nici un chip.

TABELUL 132

Veniturile vămilor în România (1881 - 1910)

Anii	Mii lei	Suma mobilă (5 termeni)	Media mobilă (5 termeni)	Media mobilă (9 termeni)	Raportul dintre (2) şi (5)	Componenta reziduală
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1881	15.649	.	.	.	.	.
1882	16.161	.	.	.	.	.
1883	20.076	84.728	16.946	.	.	.
1884	15.939	87.096	17.419	.	.	.
1885	16.903	92.137	18.427	18.759	90,1	-9,9
1886	18.017	94.024	18.805	19.666	95,8	-4,2
1887	21.202	101.010	20.202	20.124	100,4	+0,4
1888	21.963	107.917	21.583	22.239	98,8	-1,2
1889	22.925	119.177	23.835	24.593	93,2	-6,8
1890	23.810	128.090	25.618	26.063	91,4	-8,6
1891	29.277	143.255	28.651	27.501	106,5	+6,5
1892	30.116	150.464	30.093	28.875	104,3	+4,3
1893	37.127	157.609	31.522	29.954	123,9	+23,9
1894	30.134	161.901	32.380	31.375	96,0	-4,0
1895	30.955	163.457	32.961	31.211	99,2	-0,8
1896	33.569	162.042	32.408	29.936	112,1	+12,1
1897	31.672	154.249	30.830	29.390	107,8	+7,8
1898	35.712	141.096	28.219	28.183	126,7	+26,7
1899	22.341	132.725	26.545	27.597	81,0	-19,0
1900	17.802	127.315	25.463	26.757	66,5	-33,5
1901	25.198	116.463	23.293	27.066	93,1	-6,9
1902	26.262	117.517	23.503	28.012	93,8	-6,2
1903	24.860	136.065	27.213	29.558	84,1	-15,9
1904	23.395	151.054	30.211	32.571	71,8	-28,2
1905	36.350	174.418	34.884	36.002	101,0	+1,0

Scanned with OKEN Scanner

(continuare tab. 132)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
1906	40.187	199.017	39.803	39.659	101,3	+1,3
1907	49.626	224.301	44.860	.	.	.
1908	49.459	246.065	49.213	.	.	.
1909	48.679	.	.	.	.	.
1910	58.114	.	.	.	.	.

Datele extrase din ***Anuarul statistic al României****, 1915 - 1916, p. 256.*

Al doilea neajuns mai important din punct de vedere practic este faptul că această metodă nu ne permite evaluarea termenilor tendinței de lungă durată pentru o parte din primii și ultimii ani ai seriei considerate. În cazul unei medii mobile pe 9 ani, tendința de lungă durată nu poate fi determinată pentru 4 ani de la începutul și sfârșitul seriei. Cu atât mai mult, ne va fi imposibil să cunoaștem tendința de lungă durată pentru alți ani decât cei cuprinși în perioada de observație, adică previziunea, pe această cale, va fi de la început eliminată.

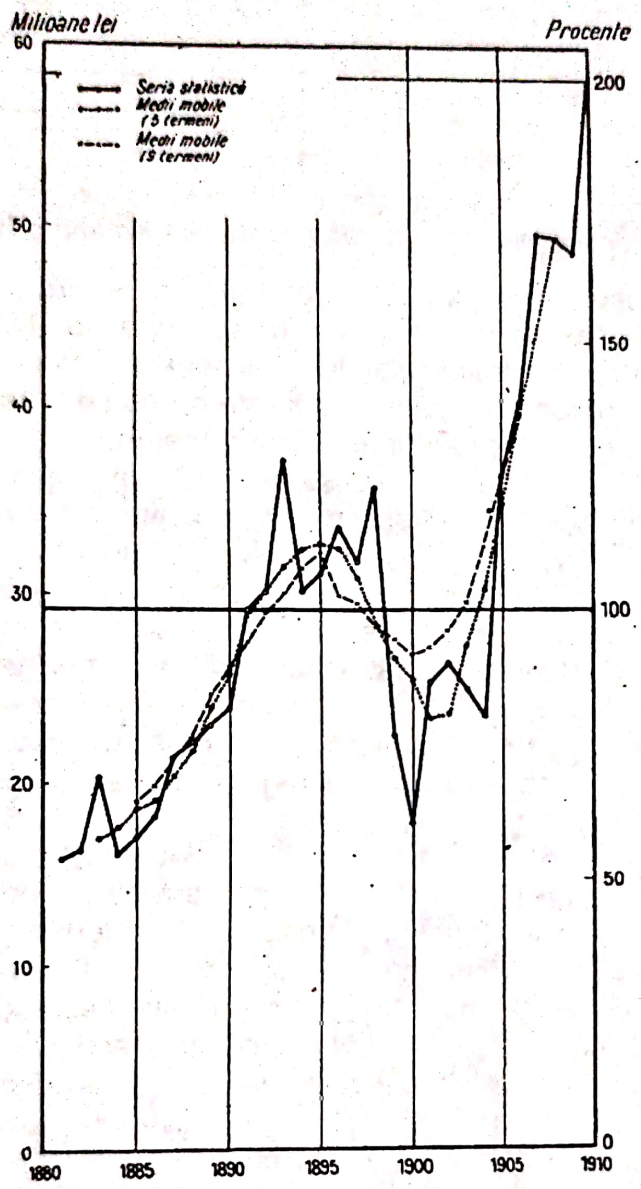
Avantajul pe care îl prezintă însă metoda mediilor mobile nu este deloc de neglijat: prin acest procedeu avem posibilitatea de a determina pe cale numerică tendința de lungă durată astfel încât aceasta să fie cât mai de aproape adaptată tuturor variațiilor capricioase ce se pot întâlni în evoluția unui fenomen. Acesta este în special cazul seriei din tabelul 132, a cărui tendință de lungă durată este din cele mai bizare și deci mai greu de determinat pe altă cale.

Din punct de vedere analitic, mediile mobile tind să niveleze orice ondulație; acesta este de fapt mecanismul pe care se bazează întrebuințarea lor. Nu trebuie totuși neglijat faptul că în cazul în care tendința de lungă durată nu ar fi reprezentată printr-o linie dreaptă, ci printr-o linie curbă, mediile mobile o vor prezenta cu o alură mult mai puțin curbată decât o are aceasta în realitate. Lucrul acesta poate să fie verificat într-o anumită măsură în cazul tabelului 132. Într-adevăr, pe graficul corespunzător (graficul 61)[1] constatăm că linia reprezentând tendința de lungă durată, calculată pe baza mediilor mobile de 9 termeni, se găsește întotdeauna în interiorul concavității liniei calculate pe

[1] *Graficul acesta este întocmit în conformitate cu principiul propus de* dr. J. Bertillon *și admis de către Institutul Internațional de Statistică (§ 28). Adică s-a adoptat pentru ordonate o astfel de unitate de măsură încât media aritmetică a seriei să fie reprezentată grafic printr-o lungime egală cu aceea care corespunde unui interval de 30 de ani pe scara timpului. Aplicarea acestui principiu este mult ușurată dacă se reduc în prealabil termenii seriei în procente față de medie și dacă se recurge la o scară de procente. Trebuie să recunoaștem totuși că principiul lui Bertillon nu conduce întotdeauna la grafice prea elocvente și nici prea estetice. Din această cauză, am renunțat la aplicarea lui constantă pentru alcătuirea graficelor din această carte.*

Scanned with OKEN Scanner

baza mediilor mobile de 5 termeni. Curbura celei dintâi va fi, prin urmare, mult mai mică decât a celei de-a doua linii.[1]



Tabelul 132

Graficul 61. Veniturile vămilor din România.

[1] *În ceea ce privește aplicarea efectivă a acestei metode, este indicat ca în primul rând să calculăm totalurile mobile (tabelul 132, coloana 3). Acestea se obțin unul din altul, prin scăderea unui termen și adăugarea altuia. Operația poate să fie făcută în mod recurent pe o mașină de calculat. Pentru a urmări mai ușor calculele, întrebuințarea unui carton în care s-au practicat două tăieturi la distanța celor doi termeni care intervin în obținerea unui total din cel precedent ne poate înlesni mult această operație, destul de anevoioasă de altfel.*

Scanned with OKEN Scanner

172. *Metoda interpolării.* Cea din urmă metodă pentru determinarea tendinţei de lungă durată este bazată complet pe calculul numeric şi poartă numele de interpolare.

Problema interpolării se pune în felul următor: vom considera punctele extremităţi ale ordonatelor care conduc la historiograma seriei. Alegându-ne un tip anumit de curbă - o linie dreaptă, o parabolă etc. - vom căuta să determinăm pe aceea care va fi *cea mai apropiată* de acest ansamblu de puncte.

Soluţia problemei interpolării astfel pusă depinde în primul rând de ceea ce vom înţelege prin depărtarea de la un ansamblu de puncte la o curbă, noţiune care poate fi definită în diferite moduri. Vom menţiona pe două dintre cele mai uzitate şi, în acest scop, să ne referim la graficul 62. Depărtarea dintre dreapta *YY'* şi ansamblul de puncte *A, B, ..., O* poate fi definită fie ca suma distanţelor *Aa, Bb, ..., Oo*, fie ca suma *pătratelor* acestora: $\overline{Aa}^2 + \overline{Bb}^2 \ldots + \overline{Oo}^2$. Dreapta interpolată va fi cu totul alta, după cum rămânem la prima sau la ultima din aceste definiţii. În primul caz, dreapta interpolată va fi aceea pentru care $\sum \overline{Aa}$ va fi un minimum, în cel de-al doilea, soluţia problemei va fi cea pentru care suma pătratelor distanţelor de la puncte la dreaptă (măsurate pe ordonate) să fie cea mai mică posibilă. Primul procedeu este relativ recent, în timp ce celălalt, cunoscut sub numele de *metoda celor mai mici pătrate*, trebuie considerat ca o metodă clasică.

Vom lăsa deoparte deocamdată orice altă consideraţie asupra determinării tendinţei de lungă durată prin interpolare şi vom trece la expunerea câtorva aplicaţii concrete.[1]

173. *Interpolarea unei linii drepte prin metoda celor mai mici pătrate.* Unul din cele mai simple procedee de interpolare prin metoda celor mai mici pătrate este acela al liniei drepte. Acest caz este şi cel mai des întâlnit. Uneori interpolarea unei linii drepte pentru determinarea tendinţei de lungă durată este motivată de însăşi forma evoluţiei fenomenului considerat, alteori însă, simplitatea procedeului este singurul criteriu care determină aplicarea lui. Bineînţeles că ultimul punct de vedere nu este numai şubred, ci direct periculos pentru reuşita studiului întreprins.

Formula căutată va fi de forma:

$$Y = a + bt \qquad (142)$$

---

[1] *Rolul pe care îl joacă linia de lungă durată faţă de historiograma efectivă este, din punct de vedere formal, identic cu cel al curbei de frecvenţe faţă de poligonul frecvenţelor. Identitatea este însă numai formală. În timp ce pentru forma curbei frecvenţelor avem indicaţii apriorice - nimeni nu se va gândi să ajusteze o altă curbă decât curba normală pentru cazul frecvenţelor feţei « marca » în aruncarea banului - pentru tendinţa de lungă durată a aceluiaşi fenomen putem adopta foarte bine formule din cele mai variate din punct de vedere analitic. Există, prin urmare, o diferenţă fundamentală între bazele operaţiei de ajustare şi ale celei de interpolare. De aici şi nevoia întrebuinţării a doi termeni diferiţi.*

Scanned with OKEN Scanner

unde t reprezintă timpul, iar Y ordonata tendinței de lungă durată calculată pe baza punctului de vedere adoptat.

TABELUL 133

Consumul zahărului în România (1899 - 1913)

Anii	Consumul (mii tone) y_i	Timpul t	Produsul ty_i	t^2	Tendința Y	$100y_i/Y$	Componenta reziduală
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1899	16,4	-7	-114,8	49	14,2	115,5	+15,5
1900	15,3	-6	-91,8	36	15,7	97,5	-2,5
1901	17,1	-5	-85,5	25	17,2	99,4	-0,6
1902	19,1	-4	-76,4	16	18,7	102,1	+2,1
1903	18,8	-3	-56,4	9	20,2	93,1	-6,9
1904	19,1	-2	-38,2	4	21,7	88,0	-12,0
1905	22,1	-1	-22,1	1	23,2	95,3	-4,7
1906	26,3	0	**-485,2**	.	24,7	106,5	+6,5
1907	29,2	+1	+29,2	1	26,2	111,5	+11,5
1908	26,0	+2	+52,0	4	27,7	93,9	-6,1
1909	27,5	+3	+82,5	9	29,2	94,2	-5,8
1910	29,9	+4	+119,6	16	30,7	97,4	-2,6
1911	33,2	+5	+166,0	25	32,2	103,1	+3,1
1912	34,9	+6	+209,4	36	33,7	103,6	+3,6
1913	35,5	+7	+248,5	49	35,2	100,9	+0,9
	370,4		**+907,2**	**280**			

Date extrase din ***Anuarul statistic al României****, 1915, 1916, p. 208.*

Să considerăm, în primul rând, o serie în timp formată dintr-un număr impar de termeni[1] (tabelul 133). Alegând pentru originea timpului anul median (1 iulie 1906), valorile lui t sunt cele din coloana (3). Observăm că prin această alegere a originii, din cauza simetriei, suma valorilor lui t este nulă. *Sub această condiție*, parametrii ecuației tendinței de lungă durată sunt determinați de relațiile:

$$a = \frac{\sum y}{n} \qquad a = \frac{\sum ty}{\sum t^2} \tag{143}$$

[1] *Pentru simplitatea operațiilor și mai ales pentru a elimina variația sezonală, calculul tendinței de lungă durată se face de obicei pe baza* datelor anuale.

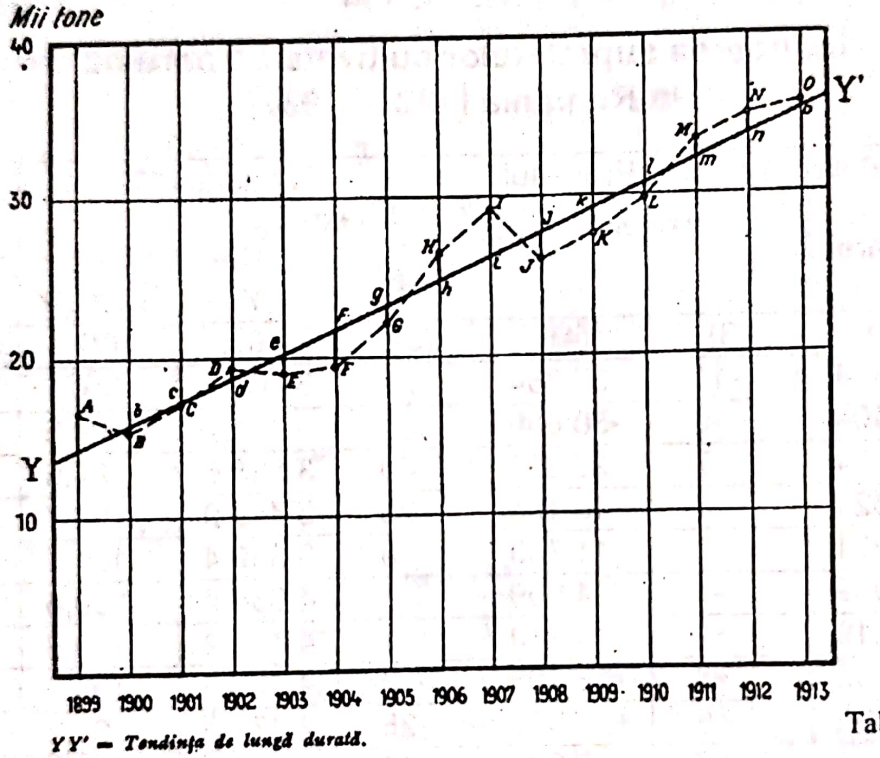
unde *y* reprezintă termenii seriei efective. Coloanele (4) și (5) ne conduc la valorile care ne permit calculul lui *b*; vom avea:

$$a = \frac{370,4}{15} = 24,7 \qquad b = \frac{907,2 - 485,2}{280} = 1,5$$

și, prin urmare, tendința de lungă durată a fenomenului considerat va fi dată de ecuația:

$$Y = 24,7 + 1,5t$$

timpul fiind raportat la originea aleasă mai sus. Coloana (6) ne înfățișează valorile datorate acestei mișcări.



YY' — Tendința de lungă durată.

Tabelul 133

Graficul 62. Tendința de lungă durată a consumului zahărului în România (1899 - 1913).

Dacă numărul termenilor seriei este par, atunci formulele (143) rămân valabile, calculul se conduce la fel, intervine însă o schimbare de amănunt. Să ne referim la tabelul 134.

Cum în acest caz nu mai există un an median, pentru atingerea scopului nostru ($\sum t = 0$),[1] vom adopta ca origine a timpului momentul care desparte

[1] *Evident că operația de interpolare se poate executa și fără ca această condiție să fie îndeplinită. Atunci însă, formulele (143) trebuie înlocuite prin altele mai complicate, ceea ce este, de fapt, în dezavantajul calculatorului.*

anul al 6-lea de al 7-lea, adică 1 ianuarie 1927. În acest caz, mijlocul anului 1927 se va găsi la distanță de 6 luni de origine, cel al anului 1928, la 18 luni etc. Adoptând ca unitate de măsură intervalul de timp egal cu 6 luni, vom obține pentru t valorile din coloana (3). Operând ca mai sus:

$$a = \frac{49.290}{12} = 4107{,}5 \qquad b = \frac{167.879 - 126.305}{572} = 72{,}7$$

și, prin urmare, tendința de lungă durată va fi dată de ecuația:

$$Y = 4107{,}5 + 72{,}7t.$$

TABELUL 134

Întinderea suprafețelor cultivate cu porumb în România (1921 - 1932)

Anii	Suprafața (mii hectare) y_i	Timpul t	Produsul ty_i	t^2	Tendința Y	$100y_i/Y$	Componenta reziduală (7) - 100
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1921	3 444	-11	-37 884	121	3 307,8	104,1	+4,1
1922	3 404	-9	-30 636	81	3 453,2	98,6	-1,4
1923	3 404	-7	-23 828	49	3 598,6	94,6	-5,4
1924	3 621	-5	-18 105	25	3 744,0	96,7	-2,3
1925	3 931	-3	-11 793	9	3 889,4	101,1	+1,1
1926	4 059	-1	-4 059	1	4 034,8	100,6	+0,6
1927	4 219	+1	+4 219	1	4 180,2	100,9	+0,9
1928	4 455	+3	+13 365	9	4 325,6	103,0	+3,0
1929	4 795	+5	+23 975	25	4 471,0	107,2	+7,2
1930	4 427	+7	+30 989	49	4 616,7	95,9	-4,1
1931	4 755	+9	+42 795	81	4 761,8	99,9	-0,1
1932	4 776	+11	+52 536	121	4 907,2	97,3	-2,7
	49 290		**-126 305**	**572**			
			+167 879				

*Date extrase din diferitele ediții ale **Anuarului statistic al României**.*

În ceea ce privește rezultatele de mai sus, avem de făcut unele precizări. În primul rând, este de relevat faptul că între b și coeficientul de regresie liniară introdus în Capitolul XI există o analogie formală. Formula (143) este identică cu formula (95). De fapt, b poate fi privit ca un coeficient de regresie a lui y în t. Cu toate acestea, nu trebuie să uităm că problema precedentă se deosebește

fundamental de aceea a regresiei. Pe când *b* nu are decât un suport pur formal, coeficientul de regresie este o valoare aparţinând complexului care defineşte o legătură stocastică stabilă.

Coeficientul *b*, cunoscut şi sub numele de panta liniei de tendinţă, are totuşi şi el o interpretare concretă: valoarea lui reprezintă variaţia lui *y* corespunzătoare unei *unităţi de timp* şi datorată numai tendinţei evolutive a fenomenului. Din primul exemplu tratat mai sus deducem că în România, pentru perioada considerată, consumul de zahăr a crescut în medie cu 1 500 tone anual sau cu 125 tone lunar; în cel de-al doilea exemplu, valoarea lui *b* se raportează la un interval de 6 luni şi, prin urmare, suprafaţa însămânţată cu porumb a crescut în medie de la un an la altul cu 2 x 72,7 = 145,4 mii ha. Sunt aici chestiuni mărunte, dar care trebuie totuşi privite cu toată atenţia.

Operaţiile necesare determinării lui *b* pot fi în parte simplificate prin aplicarea unor formule care ne dau direct suma $\sum t^2$. Acestea sunt:

$$\sum t^2 = \frac{m(m+1)(2m+1)}{3} \quad \text{când } n = 2m+1 \qquad (144)$$

$$\sum t^2 = \frac{2m(2m-1)(2m+1)}{3} \quad \text{când } n = 2m. \qquad (144 \text{ bis})$$

Astfel, pentru tabelul 133, $n = 2 \times 7 + 1$ şi deci:

$$\sum t^2 = \frac{7 \times 8 \times 15}{3} = 280$$

iar pentru tabelul 134, $n = 2 \times 6$:

$$\sum t^2 = \frac{12 \times 11 \times 13}{3} = 572.$$

174. *Interpolarea unei drepte prin metoda celor mai mici distanţe.* Interpolarea unei linii drepte pentru tendinţa de lungă durată se poate face adoptând celălalt criteriu, căutând adică dreapta pentru care suma distanţelor de la punctele ansamblului la aceasta să fie minimă. Deşi acest criteriu se impune de la început ca mult mai logic decât cel al celor mai mici pătrate, este foarte rar adoptat în practică din cauza operaţiilor de calcul mai complicate la care dă naştere.

Pentru interpolarea unei drepte pe această cale, se admite de la început, pentru motive de simetrie, că dreapta căutată va trece prin centrul de greutate al punctelor considerate, punct ale cărui coordonate sunt, cu notaţiile de mai sus, $(0, \frac{\sum y}{n})$. Adoptând centrul de greutate al punctelor ca originea axelor de coordonate, vom determina, în primul rând, abaterile valorilor *y* faţă de medie:

Scanned with OKEN Scanner

$$\eta_i = y_i - \frac{\sum y}{n}$$

În al doilea rând, vom calcula rapoartele $\frac{\eta}{t}$ și le vom ordona în ordinea lor de mărime crescătoare (tabelul 135, coloana 2).

TABELUL 135

Interpolarea unei linii drepte prin metoda celor mai mici distanțe

Anii	$\frac{\eta}{t}$	$\lvert t \rvert$	Valori cumulate $\lvert t \rvert$
(1)	(2)	(3)	(4)
1930	+45,6	7	7
1926	+48,5	1	8
1925	+58,8	3	11
1921	+60,3	11	22
1932	+60,8	11	33
1931	**+71,9**	9	**42**
1922	+78,2	9	51
1924	+97,3	5	56
1923	+100,5	7	63
1927	+111,5	1	64
1928	+115,8	3	67
1929	+137,5	5	72

Tabelul 134

Se întocmește apoi seria cumulată a valorilor lui $\lvert t \rvert$ în aceeași ordine și se determină primul termen care întrece $\frac{1}{2}\sum \lvert t \rvert$.[1] În cazul de mai sus, $\frac{1}{2}\sum \lvert t \rvert = 36$, deci termenul căutat va fi 42. Acestuia îi corespunde valoarea raportului $\frac{\eta}{t}$ cu care este egal *b*; deci:

$$b = 71{,}9$$

și, prin urmare:

[1] *N. St. Georgescu,* ***Le problème de la recherche des composantes cycliques d'un phénomène****, Journal de la Société de Statistique de Paris, octobre 1930, p. 256-257.*

$$Y = 4107{,}5 + 71{,}9t.$$

Valoarea lui *b* obținută pe această cale nu diferă, în general, simțitor de cea determinată prin metoda celor mai mici pătrate. LUCIEN MARCH recomandă să se adopte pentru *b* valoarea mediană a șirului $\frac{\eta}{t}$ [1]. Acest punct de vedere nu mi se pare justificat decât sub anumite ipoteze suplimentare[2]. De altfel, în cazul de mai sus, ar trebui să avem:

$$b = \frac{71{,}9 + 78{,}2}{2} = 75{,}1$$

valoare care nu poate fi comparabilă cu cele obținute deja.

175. *Coeficientul de liniaritate.* Calculele descrise în paragrafele precedente pot fi aplicate oricărei serii în timp. Aceasta nu înseamnă că tendința seculară a oricărei serii statistice poate fi verosimil considerată ca reprezentată de o linie dreaptă. Natural că și în această direcție există o gradație: alura generală a unei historiograme poate fi mai apropiată sau mai depărtată de linia dreaptă. S-au construit și indici care să concretizeze printr-o valoare numerică *gradul de liniaritate* al unei serii în timp. Un astfel de indice, purtând de altfel, ca și alții care se mai pot imagina, numele de *indice de liniaritate*, este dat de formula:

$$i = \frac{\sum t\eta}{\sqrt{\sum t^2 \cdot \sum \eta^2}}. \qquad (145)$$

Indicele de liniaritate variază între -1 și +1 și atinge aceste limite în cazul liniarității perfecte. Ceea ce ne interesează este mai ales valoarea absolută a acestuia. Cu cât valoarea absolută a indicelui *i* va fi mai aproape de 1, cu atât vom fi mai îndreptățiți să privim tendința de lungă durată a seriei ca reprezentată de o linie dreaptă.

Pentru cazurile deja examinate, avem:

tabelul 133: $i = \frac{422}{433{,}6} = 0{,}973$

tabelul134: $i = \frac{41574}{43277} = 0{,}961$

valori care confirmă ipotezele făcute asupra liniarității tendințelor de lungă durată ale fenomenelor considerate.

---

[1] *Lucien March, **Parallélisme, Corrélation, Causalité**, Revue de l'Institut International de Statistique, 1933, nr. 1, p. 9.*

[2] *N. St. Georgescu, **op. cit.**, p. 257.*

Scanned with OKEN Scanner

176. *Curba exponențială. Proprietăți.* Un alt tip de curbă după care se calchează în mod frecvent tendințele seculare ale anumitor fenomene este curba exponențială; ecuația acesteia este:

$$Y = Ae^{bt}. \tag{146}$$

Aplicând logaritmii în ambii membri, relația precedentă se mai poate scrie:

$$\log Y = \log A + bt \log e \tag{146 bis}$$

când apare sub o formă liniară în t. Acest fapt are o interpretare intuitivă și ne pune în evidență proprietatea esențială a curbei exponențiale, proprietate care stă la baza însăși a utilizării acesteia în statistica matematică.

Am văzut, cu prilejul examinării tendinței de lungă durată liniară, că aceasta se manifestă printr-o creștere (descreștere) constantă de la un interval la altul, mărimea acestei creșteri fiind dată de coeficientul b.

Să ne referim la tabelul 136. Reprezentând grafic atât creșterea absolută, cât și creșterea relativă a țițeiului brut întrebuințat la fabricație, constatăm că în timp ce prima are o tendință netă de sporire, cea de-a doua, deși nu matematic constantă, poate fi privită totuși ca o constantă statistică. Or, tocmai acest lucru constituie proprietatea menționată a curbei exponențiale: *constanța creșterii relative*.

O prezentare și mai intuitivă a acestui mecanism poate fi făcută cu ajutorul procesului *capitalizării*.

Să presupunem două depuneri spre fructificare de 1 leu și 5 lei, cu o dobândă de 20%. Capitalizarea se poate face fie prin dobândă simplă, fie prin dobândă compusă. În primul caz, creșterea capitalului de la un an la altul este constantă și egală cu dobânda capitalului inițial, adică de 0,20 lei și 1,00 lei respectiv. Evoluția capitalurilor este reprezentată prin două linii drepte (graficul 63 A). În sistemul dobânzii compuse, creșterea capitalului de la un an la altul nu mai este aceeași, deoarece dobânda fiecărui an, adăugându-se la capital, mărește venitul acestuia pe anul următor. Cu toate acestea, această creștere este *proporțional constantă, deoarece procentul rămâne același de-a lungul timpului.* Evoluțiile capitalurilor de 1 leu și 5 lei sunt reprezentate grafic prin două curbe *neparalele*, ultima având o tendință de ascensiune mai pronunțată decât prima (graficul 63 B). Ambele, privite însă din punctul de vedere al continuității, dau naștere la *curbe exponențiale*[1].

[1] *A. Barriol,* ***Théorie et pratique des opérations financières****, 3-e édition, Doin, Paris, 1925, p. 101.*

TABELUL 136

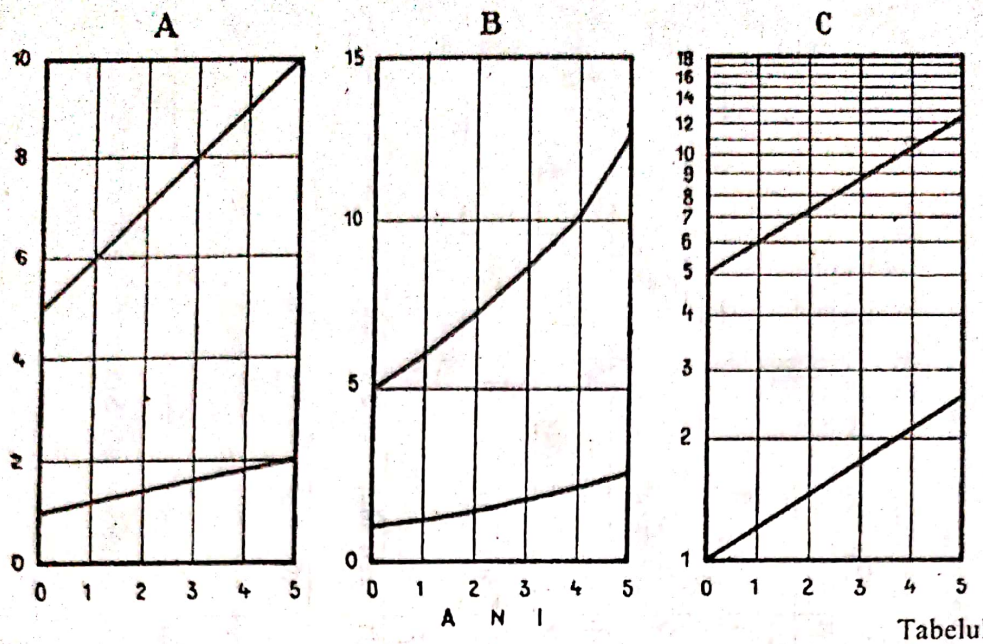
Activitatea rafinăriilor de petrol din România (1920 – 1931)

(Cantitatea de țiței brut întrebuințat la fabricație, tone)

LUNA	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931
Ianuarie	59.390	69.111	77.591	95.943	111.432	144.546	191.862	289.408	284.696	374.435	385.311	522.541
Februarie	71.718	108.627	91.692	101.437	88.972	160.347	197.087	244.959	312.428	288.981	410.842	511.319
Martie	85.997	94.840	95.752	102.425	124.823	183.016	244.695	274.412	324.953	321.513	471.307	564.033
Aprilie	55.976	78.424	105.761	115.322	112.524	164.411	223.867	258.018	315.411	361.554	457.569	574.244
Mai	96.951	81.154	120.169	143.352	179.002	187.091	231.088	322.189	259.118	399.737	483.803	582.067
Iunie	80.390	80.243	108.806	121.592	133.751	187.746	267.599	267.662	307.510	410.832	468.919	568.390
Iulie	86.553	75.613	117.975	110.998	161.375	180.215	281.298	332.603	363.626	421.508	532.968	583.994
August	91.328	80.789	83.603	114.538	130.530	167.332	292.679	321.788	405.650	442.567	501.443	561.382
Septembrie	93.111	110.345	108.746	121.123	150.981	198.766	281.342	308.044	362.071	409.237	494.816	540.735
Octombrie	88.843	67.570	105.181	119.220	143.480	183.088	295.891	307.417	355.935	402.531	451.301	541.997
Noiembrie	80.034	98.667	89.432	96.462	133.825	206.603	275.795	309.810	365.393	386.824	433.605	493.649
Decembrie	98.035	101.609	108.115	94.810	173.449	187.988	306.574	297.036	375.480	427.844	524.732	539.098
Total	**988.326**	**1.046.992**	**1.212.823**	**1.337.222**	**1.644.144**	**2.151.149**	**3.089.777**	**3.533.346**	**4.132.271**	**4.647.563**	**5.616.616**	**6.583.449**
Media lunară	**82.361**	**87.249**	**101.069**	**111.435**	**137.012**	**179.262**	**227.481**	**294.446**	**344.356**	**387.297**	**468.051**	**548.621**
Creșterea absolută	-	4.888	13.820	10.366	25.577	42.250	48.219	66.965	49.910	42.941	80.754	480.570
Creșterea relativă	-	5,9	15,8	10,3	23,0	30,8	26,9	29,4	17,0	12,5	20,9	17,2

*Date extrase din ediţiile succesive ale **Statisticii miniere a României**.*

Scanned with OKEN Scanner



Tabelul 137

Graficul 63. Comparație între scara aritmetică și scara logaritmică.

Să vedem acum ce se întâmplă dacă, în loc de a urmări variația capitalului, urmărim variația logaritmului acestuia (tabelul 137).

TABELUL 137

Capitalizare prin dobândă compusă

Anul	Capital inițial = 1 leu	Logaritmi	Capital inițial = 5 lei	Logaritmi
0	1	0	5	0,69897
1	1,2	0,07918	6	0,77815
2	1,44	0,15836	7,2	0,85733
3	1,728	0,23754	8,64	0,93651
4	2,0736	0,31672	10,368	1,01569
5	2,4832	0,39590	12,4416	1,09487

Se constată că pentru logaritmi creșterea de la un an la altul este constantă în ambele cazuri și egală cu 0,07918. Și este firesc să fie așa: fiecare termen este obținut din cel precedent, prin înmulțirea cu factorul 1,2, al cărui logaritm este numărul de mai sus[1]. Prin urmare, dacă am reprezenta grafic

[1] *Seria formată din acest fel de termeni este o progresie geometrică. Se stabilește astfel o legătură între seria geometrică și curba exponențială.*

variaţia logaritmilor termenilor unei serii exponenţiale, vom obţine două *drepte paralele*, dreapta caracterizând variaţia relativă constantă în intervale egale (graficul 63 C). Pe graficul corespunzător putem nota pe scara ordonatelor în dreptul fiecărui logaritm valoarea numărului care corespunde acestuia, în scopul de a înlesni apropierea dintre grafic şi seria efectivă. Scara astfel transformată poartă numele de scară logaritmică sau, dacă vrem să fim precişi, de scară monologaritmică, deoarece transformarea are loc numai pentru ordonate şi nu şi pentru abscise.

Întrebuinţarea acestei scări este de mare folos statisticianului. Forma curbei exponenţiale nefiind tot aşa de familiară spiritului nostru ca aceea a liniei drepte, o reprezentare grafică a unei serii în timp pe o scară logaritmică va constitui mijlocul cel mai comod prin care să ne dăm seama dacă tendinţa de lungă durată a unei serii poate sau nu poate fi verosimil primită ca reprezentată de o curbă exponenţială. Hârtie liniată pe o astfel de scară logaritmică există chiar în comerţ. O astfel de liniatură permite să reprezentăm grafic variaţia logaritmilor, fără ca în prealabil să-i fi obţinut din tabele.

Toate proprietăţile curbei exponenţiale se evidenţiază, prin urmare, mai uşor atunci când aceasta este considerată sub forma (146 bis).

177. *Interpolarea curbei exponenţiale prin metoda celor mai mici pătrate.* Pentru exemplificarea procedeului de interpolare a curbei exponenţiale, ne vom referi la datele din tabelul 136. Conform cu principiul expus mai sus, prima grijă a noastră va fi de a face o reprezentare grafică a acestei serii, pe o scară logaritmică. Historiograma astfel obţinută (graficul 64) prezintă în mod cert o tendinţă evolutivă liniară şi, în consecinţă, interpolarea unei curbe exponenţiale pentru seria efectivă este, din acest punct de vedere, motivată. Aceeaşi serie reprezentată pe o scară obişnuită ne va permite să apreciem pe o cale mai intuitivă forma curbei exponenţiale şi să ne dăm în acelaşi timp seama de concordanţa între curba interpolată şi seria statistică efectivă (graficul 65).

Calculele necesare vor avea ca suport, aşa cum am accentuat la timp, datele anuale şi nu cele lunare, iar principiul metodei va fi identic cu cel întrebuinţat pentru interpolarea unei linii drepte şi va consta din utilizarea ecuaţiei (146 bis). Amănuntele asupra conducerii calculelor se văd în tabelul 138[1].

Cu ajutorul relaţiilor:

$$\log A = \frac{\sum \log y_i}{n} \qquad b \log e = \frac{\sum t \log y_i}{\sum t^2} \tag{147}$$

[1] *Pentru a fi consecvenţi principiului de a utiliza pentru o astfel de interpolare transformarea logaritmică, ar fi trebuit ca mediile lunare să fie calculate pe baza mediei geometrice, şi nu aritmetice. Există totuşi motive care pledează pentru ultima formulă: dacă se întâmplă, de pildă, ca termenul unei singure luni să fie dintr-o cauză oarecare nul, oricare ar fi valoarea celorlalţi termeni lunari din acelaşi an, media geometrică va fi nulă. Or, este greu de conceput că valoarea medie lunară a anului corespunzător să fie şi ea nulă.*

Scanned with OKEN Scanner

a căror analogie cu (143) este motivată de forma (146 bis), obținem în cazul considerat:

$$\log A = \frac{63{,}65897}{12} = 5{,}30491 \qquad b\log e = 0{,}04066. \qquad (148)$$

Valorile acestea ne permit, sprijiniți pe relația (146 bis), să obținem $\log Y$ și apoi pe Y, adică înseși valorile datorate tendinței de lungă durată (coloanele 6 și 7).

TABELUL 138

Interpolarea curbei exponențiale pentru cantitatea de țiței brut întrebuințată în rafinăriile din România

Anii	Media lunară y_i	$\log y_i$	t	$t \log y_i$	$\log Y$	Tendința secularăă	$100y_i/Y$	Componenta reziduală
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1920	82 361	4,91573	-11	-54,07303	4,85765	72 054	114,3	+14,3
1921	87 249	4,94077	-9	-44,46693	4,93897	86 890	100,4	+0,4
1922	101 069	5,00462	-7	-35,03234	5,02029	104 783	96,5	-3,5
1923	111 435	5,04702	-5	-25,23510	5,10161	126 360	88,2	-11,8
1924	137 012	5,13676	-3	-15,41028	5,18293	152 382	89,9	-10,1
1925	179 262	5,25349	-1	-5,25349	5,26425	183 758	97,6	-2,4
1926	227 481	5,35694	+1	+5,35694	5,34557	221 600	102,7	+2,7
1927	294 446	5,46901	+3	+16,40703	5,42689	267 231	110,2	+10,2
1928	344 356	5,53701	+5	+27,68505	5,50821	322 262	106,9	+6,9
1929	387 297	5,58805	+7	+39,11635	5,58953	388 627	99,7	-0,3
1930	468 051	5,67030	+9	+51,03270	5,67085	468 656	99,9	-0,1
1931	548 621	5,73927	+11	+63,13197	5,75217	565 157	97,1	-2,9
		63,65897		**-179,47117**				
				+202,73004				

Tabelul 136

Cum $\log e = 0{,}43429$, din relațiile (148) deducem:

$$A = 201.795 \qquad b = 0{,}0936$$

și deci ecuația tendinței de lungă durată va fi:

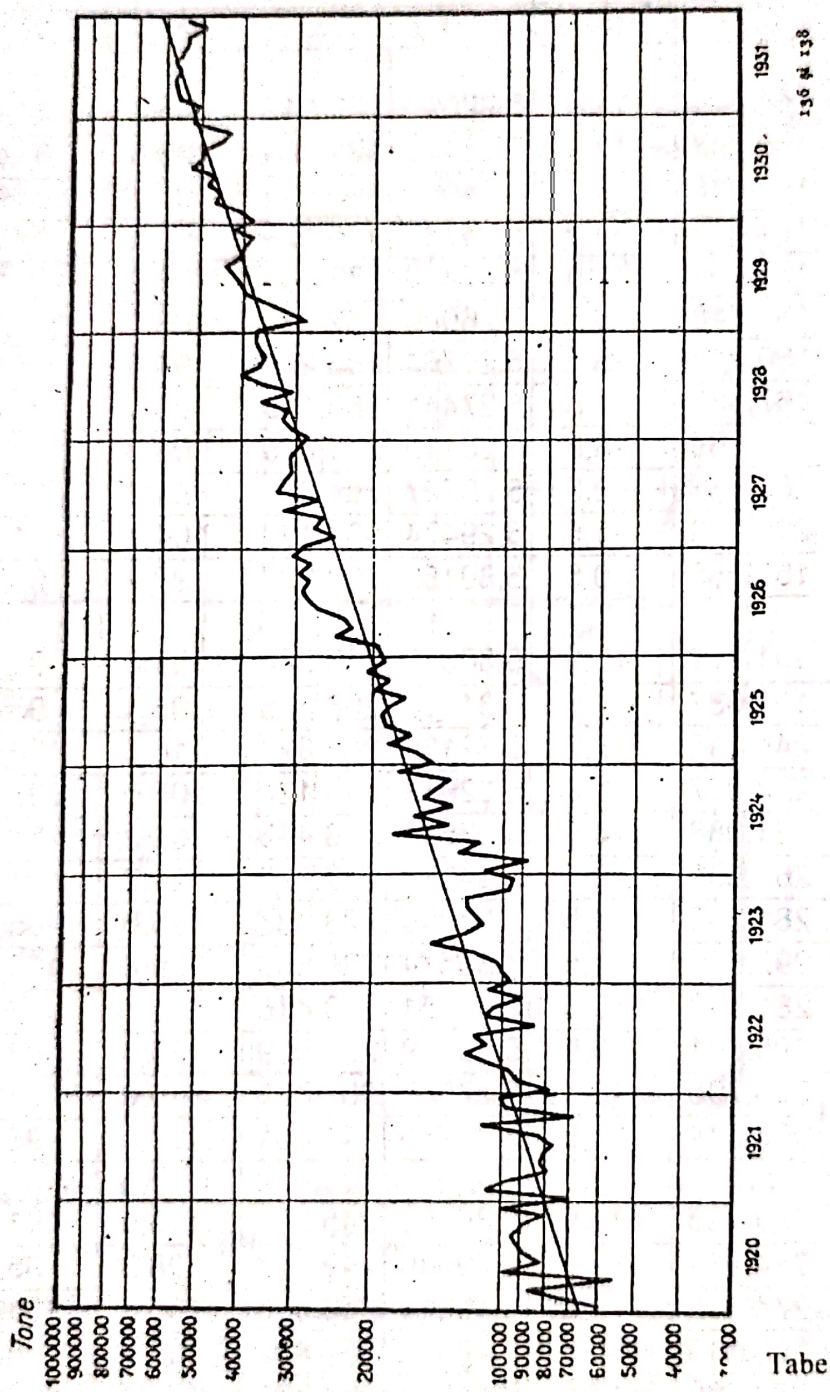
$$Y = 201.795e^{0{,}0936t}.$$

Nu trebuie uitat însă că aici timpul este măsurat în unități egale cu 6 luni. Dacă vrem acum să trecem la altă unitate de măsură pentru timp, luna, de exemplu, va trebui să împărțim prima valoare a lui b prin 6. Așa încât, adoptând unități de timp egale cu o lună, ecuația de mai sus devine:

$$Y = 201.795e^{0{,}0156t}.$$

Scanned with OKEN Scanner

Tabelul 139 ne permite o comparație între datele lunare efective și cele obținute prin interpolare. Acestea din urmă sunt obținute prin aceeași metodă ca cea întrebuințată în cazul datelor anuale, cu simpla deosebire că, în locul vechiului coeficient al lui t, vom avea de data aceasta valoarea $\frac{0{,}04066}{6} = 0{,}006777$.



Tabelele 136 și 138

Graficul 64. Tendința de lungă durată a activității rafinăriilor de petrol din România (1920 - 1931).

Scanned with OKEN Scanner

TABELUL 139

Activitatea rafinăriilor de petrol din România

LUNA	Consumul de țiței y_i	Timpul în luni t	log Y	Tendința seculară Y	$100y_i/Y$	Variația sezonală	Componenta reziduală
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1925							
Ianuarie	144.546	-11,5	5,22697	168.642	85,7	91,4	-5,7
Februarie	160.347	-10,5	5,23375	171.296	93,6	95,3	-1,7
Martie	183.016	-9,5	5,24053	173.992	105,2	104,6	+0,6
Aprilie	164.411	-8,5	5,24731	176.729	93,0	98,2	-5,2
Mai	187.091	-7,5	5,25408	179.508	104,2	109,8	-5,6
Iunie	187.746	-6,5	5,26086	182.330	103,0	101,1	+2,9
Iulie	180.215	-5,5	5,26764	185.200	97,3	105,6	-8,3
August	167.332	-4,5	5,27441	188.109	89,0	103,8	-14,8
Septembrie	198.766	-3,5	5,28119	191.070	104,0	102,3	+1,7
Octombrie	183.088	-2,5	5,28797	194.074	94,3	98,1	-3,8
Noiembrie	206.603	-1,5	5,29474	197.123	104,8	91,3	+13,5
Decembrie	187.988	-0,5	5,30152	200.227	98,9	98,5	+0,4
1926							
Ianuarie	191.862	+0,5	5,30830	203.376	94,3	91,4	+2,9
Februarie	197.087	+1,5	5,31508	206.576	95,4	95,3	+0,1
Martie	244.695	+2,5	5,32185	209.820	116,6	104,6	+12,0
Aprilie	223.867	+3,5	5,32863	213.124	105,0	98,2	+6,8
Mai	231.088	+4,5	5,33541	216.475	106,8	109,8	-3,0
Iunie	267.599	+5,5	5,34218	219.875	121,7	101,1	+20,6
Iulie	281.298	+6,5	5,34896	223.337	126,0	105,6	+20,4
August	292.679	+7,5	5,35574	226.853	129,0	103,8	+25,2
Septembrie	281.342	+8,5	5,36251	230.416	122,1	102,3	+19,8
Octombrie	295.891	+9,5	5,36929	234.039	126,4	98,1	+28,3
Noiembrie	275.795	+10,5	5,37607	237.722	116,0	91,3	+24,7
Decembrie	306.574	+11,5	5,38285	241.461	127,0	98,5	+28,5
1927							
Ianuarie	289.408	+12,5	5,38962	245.256	118,0	91,4	+26,6
Februarie	244.959	+13,5	5,39640	249.117	98,3	95,3	+3,0
Martie	274.412	+14,5	5,40318	253.035	108,4	104,6	+3,8
Aprilie	258.018	+15,5	5,40995	257.012	100,4	98,2	+2,2
Mai	322.189	+16,5	5,41673	261.053	123,4	109,8	+13,6
Iunie	267.662	+17,5	5,42351	265.163	100,9	101,1	-0,2
Iulie	332.603	+18,5	5,43028	269.325	123,5	105,6	+17,9

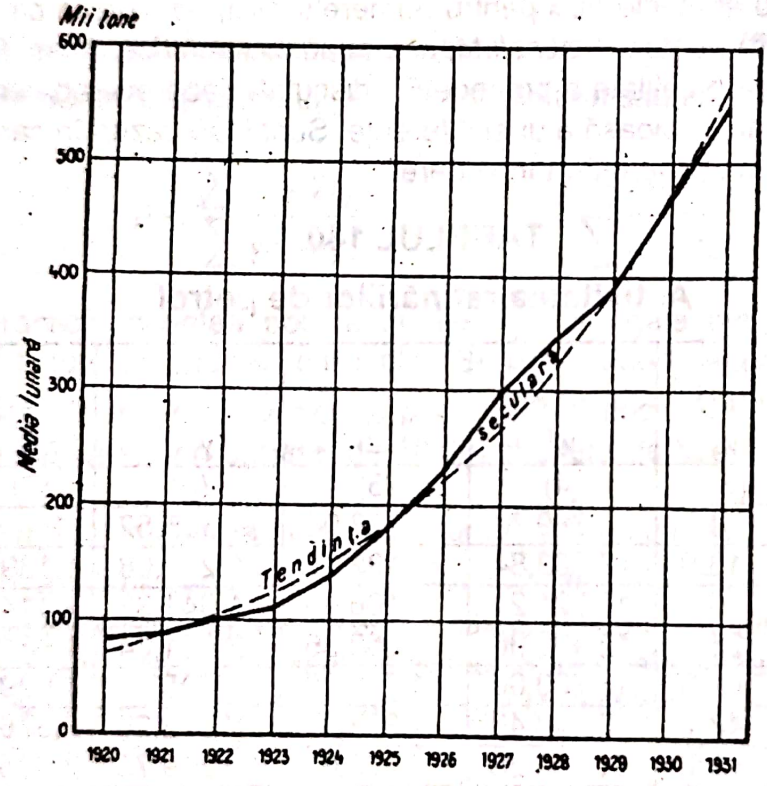
Scanned with OKEN Scanner

(continuare tab. 139)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
August	321.788	+19,5	5,43706	273.563	117,6	103,8	+13,8
Septembrie	308.044	+20,5	5,44384	277.869	110,9	102,3	+8,6
Octombrie	307.417	+21,5	5,45062	282.240	108,9	98,1	+10,8
Noiembrie	309.810	+22,5	5,45739	286.673	108,1	91,3	+16,8
Decembrie	297.036	+23,5	5,46417	291.187	102,0	98,5	+3,5

Tabelul 136

Cum momentul ales ca origine a timpului a fost 1 ianuarie 1926, pentru a determina valorile lui *t* corespunzătoare diferitelor luni, vom observa că de la origine şi până la mijlocul lunii ianuarie 1926 este o jumătate de lună, deci valoarea lui *t* va fi în acest caz 0,5. Celelalte valori se stabilesc apoi în mod automat, adăugând sau scăzând, după caz, câte o unitate.



Tabelul 138

Graficul 65. Evoluţie exponenţială reprezentată pe scara aritmetică.

178. *Creşterile relative şi evoluţia exponenţială.* Fiecare termen al seriei valorilor tendinţei de lungă durată provine din precedentul prin înmulţirea cu aceeaşi *raţie.* Determinarea raţiei anuale se face căutând numărul care corespunde logaritmului 2 x 0,04066. Acesta este 1,206, de unde deducem valoarea creşterii relative medii a fenomenului considerat: 20,6%.

Scanned with OKEN Scanner

Aceeași problemă se poate pune și pentru creșterea relativă *lunară*. Vom avea:

$$e^{b'} = 1{,}0157.$$

Creșterea relativă lunară va fi 1,57% și nu poate fi, prin urmare, dedusă, în cazul unei curbe exponențiale, prin simpla împărțire a creșterii anuale cu 12.[1]

179. *Alt procedeu de interpolare a unei curbe exponențiale.* Procedeul interpolării unei curbe exponențiale, expus în paragrafele precedente, este bazat pe principiul celor mai mici pătrate, aplicat însă logaritmilor datelor, și nu acestora înșeși. Or, nu este greu să ne dăm seama, fie chiar printr-o simplă inspecție a tablelor de logaritmi, că aceeași variație a logaritmului determină variații absolute mai mari în cazul numerelor mari și mai mici în caz contrar. Sistemul conduce, prin urmare, la o echivalare a abaterilor mari pentru numerele mari cu abaterile mici pentru numerele mici, adică face ca importanța să cadă nu asupra *abaterilor absolute,* ci a *abaterilor relative*. Acest fapt trebuie privit în general ca o calitate a procedeului discutat, deoarece abaterea relativă este o măsură mai judicioasă a unei diferențe. Sunt însă cazuri în care abaterea absolută trebuie avută neapărat în vedere.

TABELUL 140

Activitatea rafinăriilor de petrol

Anii	Media lunară y_i	Diferența $\Delta_i = y_{i+1} - y_i$	y_i^2	Δ_i^2	$y_i \Delta_i$	x_i	Tendința Ax_i
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1920	82,4	4,8	6 789,76	23,04	395,52	1,	80,5
1921	87,2	13,9	7 603,84	193,21	1 212,08	1,189	95,7
1922	101,1	10,3	10 221,21	106,09	1 041,33	1,414	113,8
1923	111,4	25,6	12 409,96	655,36	2 851,84	1,681	135,3
1924	137,0	42,3	18 769,00	1 789,29	5 795,10	1,999	160,9
1925	179,3	48,2	32 148,49	2 323,24	8 642,26	2,376	191,3
1926	227,5	66,9	51 756,25	4 475,61	15 219,75	2,825	227,4
1927	294,4	50,0	86 671,36	2 500,00	14 720,00	3,359	270,4
1928	344,4	42,9	118 611,36	1 840,41	14 774,76	3,994	321,5
1929	387,3	80,8	150 001,29	6 528,64	31 293,84	4,749	382,3
1930	468,1	80,5	219 117,61	6 480,25	37 682,05	5,647	454,6
1931	548,6	-	-	-	-	6,714	540,5
			714 100,13	**26 915,14**	**133 628,53**		

Tabelul 136

[1] *A se vedea Cap. IV, § 59-60.*

Pentru aceasta, procedeul de interpolare a unei curbe exponenţiale diferă de precedentul şi este, în acelaşi timp, şi mai laborios.

Se calculează, în primul rând, diferenţele succesive de la un termen al seriei la celălalt: $\Delta_i = y_{i+1} - y_i$, iar cu ajutorul acestor valori se obţin sumele:

$$\sum y_i^2 = 714100{,}13 \qquad \sum \Delta_i^2 = 26915{,}14 \qquad \sum y_i \Delta_i = 133628{,}53$$

(tabelul 140, coloanele 3, 4, 5, 6).

Apoi se rezolvă ecuaţia în r:

$$(\sum y^2 - r)(\sum \Delta^2 - 2r) - (\sum y\Delta + r)^2 = 0 \qquad (149)$$

care, în cazul nostru, este:

$$r^2 - 1.722.372{,}46r + 1.363.520.943{,}0073 = 0.$$

Cea mai mică rădăcină a acestei ecuaţii, $r_0 = 792{,}02$, este apoi întrebuinţată pentru obţinerea lui $p = e^b - 1$ din una sau alta din relaţiile[1]:

$$p = \frac{\sum \Delta^2 - 2r_0}{\sum y\Delta + r_o} \qquad p = \frac{\sum y\Delta + r_0}{\sum y^2 - r_o}$$

În practică, acestea pot să nu fie riguros egale din cauza diferitelor aproximaţii făcute de-a lungul calculelor. Se poate adopta, în acest caz, media lor aritmetică. Avem în cazul nostru $p = 0{,}189$ şi deci $e^b = 1{,}189$. Creşterea relativă anuală este, prin urmare, de 18,9% şi diferă de cea stabilită prin prima cale.

Pentru determinarea parametrului A din ecuaţia exponenţială se formează în prealabil seria $1, e^b, e^{2b}, e^{3b}, \ldots\ldots e^{(n-1)b}$,[2] ai cărei termeni îi vom însemna prin x (coloana 7). Operaţia se reduce la una de interpolare liniară: $Y = Ax$, care se poate efectua fie prin metoda celor mai mici pătrate, fie prin cea a distanţelor minime. Prima metodă ne conduce la relaţia de formă cunoscută:

$$A = \frac{\sum xy}{\sum x^2}.$$

Printr-un calcul, pe care nu îl mai publicăm, obţinem: $\sum xy = 12389{,}24$. Numitorul formulei precedente se poate obţine fie direct, fie prin intermediul formulei:

---

[1] *N. St. Georgescu, **op. cit.**, p. 218-225.*

[2] *Acest calcul se face mai uşor cu ajutorul logaritmilor.*

Scanned with OKEN Scanner

$$\sum x^2 = 1 + e^{2b} + e^{4b} + \ldots + e^{2(n-1)b} = \frac{e^{2nb} - 1}{e^{2b} - 1}$$

unde n este numărul termenilor. Astfel obținem:

$$\sum x^2 = \frac{62{,}727}{0{,}414} = 153{,}93 \qquad A = \frac{12389{,}24}{153{,}93} = 80{,}5$$

și, prin urmare, ecuația tendinței de lungă durată va fi:

$$Y = 80{,}5e^{0{,}173t}.$$

Aici însă originea timpului este 1 iulie 1920, iar unitatea de măsură este anul. (Cantitatea de țiței este și ea măsurată în mii de tone.) Valorile tendinței se obțin înmulțind pe x_i cu A (coloana 8).

180. *Considerații asupra metodelor de interpolare*. Metode și formule din cele mai variate sunt întrebuințate pentru interpolarea tendinței de lungă durată. Acestea ar putea face singure obiectul unui tratat. Ne oprim aici, procedeele discutate mai sus fiind suficiente pentru a ne face o idee completă despre felul cum trebuie privită operația de interpolare și pentru a ne da, în același timp, putința să facem față în mai toate problemele similare de ordin practic.

Interpolarea ca metodă de lucru prezintă, întocmai ca și celelalte metode discutate mai sus, neajunsuri și avantaje în același timp.

Deși utilizează efectiv calculul numeric, interpolarea nu este un procedeu căruia să nu i se poată imputa rezultate subiective. Primul lucru pe care trebuie să-l facem înainte de a trece la aplicarea metodei este alegerea *tipului* de curbă pe care îl vom interpola. Or, este evident că historiograma neavând un contur prea rigid, acolo unde un statistician interpolează o dreaptă, altul poate să interpoleze o parabolă, un al treilea o curbă exponențială etc. Elasticitatea alegerii tipului de curbă poate să fie totuși în parte limitată dacă această operație este precedată de o analiză *structurală* a șirului de date statistice. Această analiză nu este greu de făcut și constă, în general, dintr-o verificare în ansamblu pe seria în timp a proprietăților caracteristice ale curbelor mai des întrebuințate în statistică. Pentru cazul unei drepte, de pildă, vom căuta să controlăm realitatea constanței creșterii (descreșterii) absolute de la un termen la altul; pentru curba exponențială, acest lucru se va face pentru creșterea relativă etc.

Unde însă subiectivismul nu mai poate fi înlăturat este în ceea ce privește criteriul după care se face interpolarea și care poate fi al celor mai mici distanțe, al celor mai mici pătrate sau un altul oarecare din care s-au imaginat, dar care nu sunt expuse în lucrarea de față; acestea, la rândul lor, pot fi aplicate fie direct datelor, fie logaritmilor acestora, fie, în sfârșit, raporturilor sau altor combinații ce se pot face între primele. Din fericire, rezultatele obținute sunt,

Scanned with OKEN Scanner

indiferent de criteriul adoptat, sensibil egale între ele pentru același caz, mai ales dacă alegerea tipului de curbă s-a făcut în urma unei complete analize a structurii seriei în timp.

Dintre avantajele interpolării, trebuie să menționăm în primul rând putința de a obține valori și în afara intervalului de observație, o extrapolare, cu alte cuvinte. Acest fapt va constitui un prim pas în direcția previziunii. Spunem un prim pas, căci nu trebuie uitat că extrapolarea și previziunea sunt două lucruri cu prea puține puncte de atingere. Fenomenele nu pot ține seamă de rigiditatea formulelor noastre matematice și, după cum spuneam mai sus, evoluția lor poate fi modificată de un ansamblu foarte variat de factori. A uita, o singură clipă măcar, că interpolarea, oricât de perfectă ar fi, nu are decât o însemnătate formală și numai pentru perioada de timp observată constituie greșeala cea mai de neiertat a unui statistician.

S-a spus, pe drept cuvânt, că cea mai bună interpolare poate conduce în majoritatea cazurilor la cea mai absurdă extrapolare. Referindu-ne la ecuația tendinței de lungă durată a suprafețelor însămânțate cu porumb (§173), deducem, printr-o aplicație rigidă a ecuației obținute acolo, că în interval de 50 de ani suprafața semănată cu porumb va întrece suprafața arabilă actuală a României. Nu este greu de văzut cât de absurdă, privită din punctul de vedere al unei situații normale, este o atare previziune[1].

Dar chiar în ceea ce privește o extrapolare mai apropiată, previziunea pe această cale este cât se poate de șubredă. Ecuația tendinței de lungă durată a consumului de cărbuni în Franța (tabelul 130) este:

$$Y = 38.861 + 943\,t$$

originea timpului fiind 1 iulie 1893. În baza acestei ecuații, consumul de cărbuni ar fi pentru anul 1928 în jurul cifrei de 72 milioane tone. Or, în acest an, consumul a fost de 86.488 milioane. Diferența nu poate fi cel puțin privită ca o abatere ciclică, deoarece anul 1928 a fost mai degrabă un an de consum slab în comparație cu circumstanțele conjuncturale ale epocii.

181. *Variații sezonale. Metoda periodogramei.* Pentru determinarea nivelului pe care trebuie să-l atingă manifestarea unui fenomen, în diferitele epoci ale anului, se întrebuințează, ca și în cazul determinării tendinței de lungă durată, metode variate și complet diferite ca mecanism. Le vom menționa pe cele care sunt în general admise de către toți statisticienii.

Cea mai simplă metodă de a pune în evidență variațiile sezonale, simplă și din punctul de vedere al concepției, și din acela al procedeului de calcul, este cunoscută sub numele de *metoda periodogramei*. Termenul este motivat de faptul că variațiile sezonale, atunci când există efectiv, constituie netăgăduit o mișcare periodică. Pentru descoperirea acesteia, mediile aritmetice ale datelor

[1] *Dacă, dimpotrivă, am merge în sensul celălalt, ecuația considerată ne conduce la alt rezultat absurd: acum 30 de ani nu se semăna porumb în România.*

Scanned with OKEN Scanner

fiecărei luni în parte vor constitui elementul cel mai la îndemână. Cu toate acestea, dacă media lunii ianuarie, de pildă, este mai mică decât aceea a lunii noiembrie, nu va însemna numaidecât că variația sezonală va fi de la o lună la alta indicată de această diferență; luna noiembrie se găsește întotdeauna la un interval de 10 luni după prima, în care interval mișcarea fenomenului este influențată de tendința de lungă durată, așa că diferența considerată poate să fie o consecință a acestui fapt. Vom elimina influența tendinței de lungă durată dacă vom raporta mediile lunare la valorile tendinței de lungă durată calculate prin procedeul expus mai sus. Obținem astfel indicii neajustați (tabelul 141). Luând ca bază media acestor indici neajustați $\frac{1266}{12} = 105{,}5$, indicii ajustați se obțin prin raportarea primilor la această medie. Media indicilor ajustați trebuie să fie întotdeauna egală cu 100.

Nu trebuie trecut cu vederea că această metodă pentru determinarea variațiilor sezonale este dintre cele mai slabe, deoarece, fiind bazată pe mediile lunare, rezultatele pot fi oricând influențate de unele situații pasagere, cum ar fi o grevă sau o activitate excesivă în cursul unei anumite luni.

TABELUL 141

Variațiile sezonale ale activității rafinăriilor de petrol din România (metoda periodogramei)

Luna	Cantitatea medie de țiței consumat	Tendința[1]	Indici neajustați	Indici ajustați
Ianuarie	217 189	217 189	100	94,8
Februarie	215 701	220 599	98	92,9
Martie	240 647	224 074	107	101,4
Aprilie	235 257	227 592	103	97,6
Mai	265 477	231 241	115	109,0
Iunie	250 287	234 803	107	101,4
Iulie	270 727	238 517	114	108,1
August	266 136	242 253	110	104,3
Septembrie	264 943	246 075	108	102,4
Octombrie	255 205	249 941	102	96,7
Noiembrie	247 508	253 872	97	91,9
Decembrie	269 564	257 868	105	99,5
			1 266	**1 200,0**

Tabelul 136

[1] *Termenii din această coloană sunt deduși fiecare din cel precedent, prin înmulțirea cu* $e^{b'} = 1{,}0157$ *(§ 178).*

182. *Metoda Macauley*. Variaţiile sezonale mai pot fi determinate şi pe alte căi, mai precise decât metoda periodogramei. Una dintre acestea este cunoscută sub numele de metoda *Federal Reserve Board* sau metoda lui F. R. MACAULEY, după numele celui care a imaginat-o. Principiul călăuzitor este în parte asemănător periodogramei: raportarea datelor lunare efective la valorile corespunzătoare ale tendinţei. Acestea trebuie astfel determinate încât să nu mai cuprindă nici o urmă de variaţie sezonală. Cel mai indicat procedeu ar fi, din cauză că perioada acoperă 12 luni (§171), să adoptăm pentru calculul tendinţei o medie mobilă de 12 luni. MACAULEY recomandă ca această medie să fie centrală la cea de-a 7-a lună din şirul celor care intră în compoziţia ei. Aceasta înseamnă însă că ne abatem de la însuşi principiul mediilor mobile. Se poate remedia acest lucru dacă în locul unei medii mobile:

$$\frac{I+F+M+\ldots+O+N+D}{12}$$

de pildă, centrată în luna iulie, vom adopta pentru aceeaşi lună valoarea:

$$\frac{\frac{1}{2}I+F+M+\ldots+N+D+\frac{1}{2}I_1}{12}$$

unde I_1 este valoarea corespunzătoare lunii ianuarie din anul următor celui considerat. Neîntrebuinţarea ultimei formule a fost determinată de lungimea calculelor la care aceasta ar da naştere. Argumentul nu mi se pare valabil, căci expresia considerată se mai poate scrie:

$$\frac{(I+F+M+..+N+D)+(F+M+A+..+N+D+I_1)}{24}$$

ceea ce înseamnă că pentru calculul ei va trebui să calculăm în primul rând o sumă mobilă de 12 termeni (la fel ca şi pentru prima formulă) din care vom lua o nouă sumă mobilă de 2 termeni (tabelul 142, coloanele 3 şi 4). Rezultatele vor fi împărţite cu 24 (în loc de 12, ca mai sus) (coloana 5).

În urmă, prin raportarea datelor efective la cele calculate, obţinem pentru fiecare lună în parte un *raport de variaţie sezonală*. Tabelul 143 cuprinde toate valorile acestui raport corespunzătoare activităţii rafinăriilor de petrol în perioada 1920 - 1931[1].

[1] *Din cauza întrebuinţării mediilor mobile, nu putem obţine valorile raportului de variaţie sezonală pentru primele 6 luni din 1920 şi ultimele 6 din 1931.*

Scanned with OKEN Scanner

TABELUL 142

Variația sezonală a activității rafinăriilor de petrol din România[1] (Metoda Macauley)

Luna	Cantitatea de țiței consumată	Suma mobilă (12 termeni)	Suma mobilă (2 termeni)	Câtul prin 24	Raportul de variație sezonală
(1)	(2)	(3)	(4)	(5)	(6)
1926		2.480.190			
Ianuarie	191.862	2.581.273	5.061.463	210.894	91,0
Februarie	197.087	2.706.620	5.287.893	220.329	89,5
Martie	244.695	2.789.196	5.495.816	228.992	106,9
Aprilie	223.867	2.901.999	5.691.195	237.133	94,4
Mai	231.088	2.971.191	5.873.190	244.716	94,4
Iunie	267.599	3.089.777	6.060.968	252.536	106,0
Iulie	281.298	3.187.323	6.277.100	261.546	107,6
August	292.679	3.235.195	6.422.518	267.605	109,4
Septembrie	281.342	3.264.912	6.500.107	270.838	104,8
Octombrie	295.891	3.399.063	6.563.975	273.499	108,2
Noiembrie	275.795	3.390.164	6.689.227	278.718	99,0
Decembrie	306.574	3.390.227	6.780.391	282.516	108,5

Tabelul 136

TABELUL 143

Rapoartele de variație sezonală corespunzătoare activității rafinăriilor de petrol din România

LUNA	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931
Ianuarie	.	71,1	80,6	89,3	93,3	89,9	91,0	101,7	90,0	101,8	88,5	99,7
Februarie	.	126,1	93,4	93,5	72,8	98,3	89,5	85,1	97,3	77,7	92,9	96,7
Martie	.	109,7	97,5	92,9	100,6	109,8	106,9	94,5	99,4	85,7	105,1	105,8
Aprilie	.	90,9	106,0	103,5	89,1	96,5	94,4	88,4	95,2	95,3	100,8	106,6
Mai	.	94,2	119,1	127,7	138,9	106,9	94,4	109,7	107,0	104,6	105,6	106,8
Iunie	.	92,1	107,9	108,6	100,0	105,1	106,0	90,8	90,2	106,7	101,1	103,7
Iulie	104,6	86,3	115,9	99,0	116,6	99,4	107,6	113,0	104,5	108,7	112,5	.
August	107,8	92,6	81,2	102,1	91,4	90,6	109,4	108,4	115,6	112,5	103,7	.
Septembrie	107,5	127,4	104,9	107,6	101,9	105,3	104,8	102,1	103,5	101,1	100,6	.
Octombrie	101,1	77,0	100,8	105,1	93,9	94,4	108,2	100,4	101,3	97,0	90,2	.
Noiembrie	90,7	109,0	84,6	84,0	86,2	104,2	99,0	99,9	102,9	91,6	85,1	.
Decembrie	112,0	108,9	100,8	81,2	109,9	92,4	108,5	94,8	104,0	99,9	101,4	.

Tabelul 136

[1] *Acest tabel nu este decât o parte din tabelul complet necesar aplicării metodei Macauley, pe care, din cauza lungimii lui, nu îl publicăm în întregime.*

Scanned with OKEN Scanner

Teoretic, ar trebui ca acest raport să fie constant pentru aceeaşi lună de-a lungul seriei. Evident că acest lucru nu poate fi întâlnit în practică. Fiecărei luni îi va corespunde, prin urmare, nu o singură valoare, ci o serie de valori. Va trebui totuşi să înlocuim seria printr-o valoare unică. Astfel pusă problema, greutatea ei nu mai constă decât în alegerea tipului de valoare medie care va reprezenta seria raporturilor de variaţie sezonală. Media geometrică trebuie exclusă de la început, deoarece ar urma ca, dacă în cursul unei anumite luni a existat o eclipsă a fenomenului considerat, să adoptăm valoarea zero pentru indicele de variaţie sezonală al lunii respective; aceasta ar constitui însă o absurditate. Ne rămân, dintre valorile tipice mai uzitate, media aritmetică şi mediana. Opiniile tuturor statisticienilor s-au pronunţat de la început în favoarea ultimei. Într-adevăr, mediana este, după cum ştim, valoarea tipică ce nu suferă influenţa valorilor excepţionale ale seriei, valori care se pot prezenta în mod frecvent mai ales în seriile referitoare la fenomenele economice şi care trebuie evident lăsate de o parte în stabilirea indicelui teoretic de variaţie sezonală. Dacă totuşi fenomenul urmează o variaţie sezonală regulată, adoptarea uneia sau alteia din aceste valori tipice nu poate conduce decât la rezultate sensibil identice.

Ceea ce trebuie însă făcut, înainte chiar de a ne pune problema determinării indicilor sezonali, este o analiză care să ne arate dacă fenomenul considerat este sau nu supus unei variaţii sezonale nete. Acest lucru se poate face cu ajutorul *tabelului de multiplă frecvenţă* (tabelul 144). În acest tabel se înseamnă, pentru fiecare lună în parte şi pe o scară de variaţie adecvată, toate valorile raportului de variaţie sezonală obţinute mai sus (tabelul 143). Numai în cazul când semnele prin care însemnăm pe un astfel de tabel prezenţa unei valori a raportului de variaţie sezonală nu sunt uniform repartizate, nici pe coloane şi nici de-a lungul acestora, putem afirma existenţa unei variaţii sezonale.

O dată această analiză făcută, se adoptă pentru valoarea indicilor sezonali fie media aritmetică, fie mediana; indicii ajustaţi se obţin din primii, reducându-i în mod proporţional (la fel ca şi în metoda periodogramei) la media 100. Indicii din tabelul 144 au fost calculaţi pornind de la valoarea medianei.

Scanned with OKEN Scanner

TABELUL 144

Frecvențele valorilor raportului de variație sezonală

Valoarea raportului	Ianuarie	Februa-rie	Martie	Aprilie	Mai	Iunie	Iulie	August	Septem-brie	Octom-brie	Noiem-brie	Decem-brie
Peste 115		x			xxx		xx	x	x			
114-												
113-							x					
112-							x	x				x
111-												
110-					†							
109-			xx		x			x			x	x
108-						x	x	x		x		xx
107-					x	x	x o	x	xx			
106-			x	xx	xx o	xx	†		†			
105-			xx		x	x			x	x		
104-					x		xx		xx o		x	x
103-				x		x o		x o	x			
102-								x	x		x	
101-	xx					x†		†	xx	xx		x o †
100-			x o †	x		x			x	xx o		x
99-	x		x				xx				xx	x
98-		x										
97-		x	x	†						x†		
96-		x		x								
95-				xx o								
94-			x	x	xx					x	†	x
93-	x	xx o †								x		

Valoarea raportului	Ianuarie	Februa-rie	Martie	Aprilie	Mai	Iunie	Iulie	August	Septem-brie	Octom-brie	Noiem-brie	Decem-brie
92-		x	x			x		x				x
91-	x							x			x O	
90-	x O †			x		xx		x		x	x	
89-	xx	x		x								
88-	x			x								
87-												
86-							x				x	
85-		x	x								x	
Sub 85-	xx	xx						x		x	xx	x
Media	90,6	93,0	100,7	97,0	110,4	101,1	106,2	101,4	106,1	97,2	94,3	101,3
Mediana	90,0	93,4	100,6	95,3	106,9	103,7	107,6	103,7	104,8	100,4	91,6	101,4
Indici ajustați	**90,0**	**93,4**	**100,6**	**95,3**	**107,0**	**103,8**	**107,7**	**103,8**	**104,9**	**100,4**	**91,6**	**101,5**

Tabelul 143

Cercurile (**O**) *indică poziția medianei, crucile* (†), *pe aceea a mediei aritmetice.*

Scanned with OKEN Scanner

183. *Metoda şirului de rapoarte de la lună la lună.* În fine, metoda cea mai răspândită pentru determinarea variaţiilor sezonale a fost imaginată de prof. WARREN PERSONS de la *Harvard Committee on Economics Research* şi este cunoscută mai ales sub numele de *metoda lanţului de rapoarte de la lună la lună*[1].

Prof. PERSONS porneşte de la ideea că, pentru a determina factorul sezonal corespunzător unei luni, mijlocul cel mai firesc este raportarea datei statistice a unei luni la cea a lunii precedente. Se pare că acest punct de vedere este cel mai logic, deoarece, prin raportul de la lună la lună, se ţine seama şi de faptul că sunt anumite circumstanţe speciale fiecărei epoci în parte, circumstanţe care nu au nimic comun cu factorul sezonal şi care, admitem, vor influenţa aproape la fel manifestările fenomenului din două sau chiar mai multe luni consecutive.

O dată aceste rapoarte calculate[2] (tabelul 145), vom forma un tabel de multiplă frecvenţă a lor, întocmai ca şi pentru metoda Macauley (tabelul 146). Pe acesta din urmă, se pot vedea şi mai precis diferenţele ce există între valorile mediei aritmetice şi ale medianei şi care schimbă uneori însuşi sensul variaţiei sezonale.

Adoptând mediana rapoartelor de la lună la lună ca valoare tipică a seriei corespunzătoare fiecărei luni, obţinem 12 rapoarte teoretice pe care le vom nota cu $c_1, c_2, \ldots, c_{12}$. Ele sunt înscrise în ultima linie a tabelului 146.

Şirul numerelor c are însă dezavantajul de a constitui nişte indici care nu au aceeaşi bază. Luând ca bază luna ianuarie, vom obţine un nou şir de indici care se deduc uşor din şirul valorilor c, în baza relaţiilor evidente[3]:

$$d_1 = \frac{IAN}{IAN} = 100$$

$$d_2 = \frac{FEB}{IAN} = c_2$$

$$d_3 = \frac{MAR}{IAN} = \frac{MAR}{FEB} \cdot \frac{FEB}{IAN} = c_3 d_2$$

$$d_4 = \frac{APR}{IAN} = \frac{APR}{MAR} \cdot \frac{MAR}{IAN} = c_4 d_3$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$d_{12} = \frac{DEC}{IAN} = \frac{DEC}{NOV} \cdot \frac{NOV}{IAN} = c_{12} d_{11}$$

---

[1] Link-relatives, *în engleză.*

[2] *Datele lunii ianuarie se vor raporta la datele lunii decembrie a anului precedent.*

[3] *În toate aceste egalităţi, rapoartele trebuie considerate ca exprimate în procente. Calculul valorilor d se face mai comod cumulând logaritmii coeficienţilor c (tabelul 174, coloanele 3 şi 4), începând cu logaritmul lui 100 şi având grijă ca de fiecare dată să scădem 2 unităţi din rezultat (consecinţă a faptului că rămânem la procente).*

Scanned with OKEN Scanner

TABELUL 145

Rapoartele de la lună la lună pentru activitatea rafinăriilor de petrol

LUNA	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929	1930	1931
Ianuarie	...	70,5	76,4	88,7	117,5	83,3	102,1	94,4	95,8	99,7	90,1	99,6
Februarie	120,8	157,2	118,2	105,7	79,8	110,9	102,7	84,6	109,7	77,2	106,6	97,9
Martie	119,9	87,3	104,4	101,0	140,3	114,1	124,2	112,0	104,0	111,3	114,7	110,3
Aprilie	65,1	82,7	110,5	112,6	90,1	89,8	91,5	94,0	97,1	112,5	97,1	101,8
Mai	173,2	103,5	113,6	124,3	159,1	113,2	103,8	124,9	113,9	110,6	105,7	101,4
Iunie	82,9	98,9	90,5	84,8	74,7	100,4	115,8	83,1	85,6	102,8	96,9	97,7
Iulie	107,7	94,2	108,4	91,3	120,7	96,0	105,1	124,3	118,2	102,6	113,7	102,7
August	105,5	106,8	70,9	103,2	80,9	92,9	104,0	96,7	111,6	105,0	94,1	96,1
Septembrie	102,0	136,6	130,1	105,7	115,7	118,8	96,1	95,7	89,3	92,5	98,7	96,4
Octombrie	95,4	61,2	96,7	98,4	95,0	92,1	105,2	99,8	98,3	98,4	91,2	100,2
Noiembrie	90,1	146,0	85,0	80,9	93,3	112,8	93,2	100,8	102,7	96,1	96,1	91,1
Decembrie	122,5	103,0	120,9	98,3	129,6	91,0	111,2	95,9	102,8	110,6	121,0	109,2

Tabelul 136

Scanned with OKEN Scanner

Operaţia de recurenţă poate fi dusă mai departe, calculând un nou indice pentru luna ianuarie:

$$d_1' = \frac{IAN}{IAN} = \frac{IAN}{DEC} \cdot \frac{DEC}{IAN} = c_1 d_{12}.$$

Această nouă valoare trebuie să fie teoretic egală cu d_1, adică cu 100, deoarece reprezintă variaţia sezonală a aceleiaşi luni a anului. Or, acest lucru nu se întâmplă în practică decât în mod cu totul excepţional. Explicaţia constă în faptul că mediana şi chiar media aritmetică nu îndeplinesc condiţia de *circularitate*, adică ultima valoare obţinută pe baza unei operaţii ca cea de mai sus nu reproduce valoarea iniţială[1]. Există opinia că acest fapt îşi găseşte explicaţia şi în influenţa tendinţei de lungă durată, care ar face ca valoarea lui ianuarie după decembrie să nu fie egală cu aceea a lui ianuarie dinaintea lui decembrie. Acest punct de vedere este însă complet greşit: dacă în locul mediei aritmetice sau medianei se adoptă o altă valoare care să îndeplinească condiţia de circularitate, cum este, de pildă, media geometrică, atunci diferenţa între d_1 şi d_1' nu se mai produce.

Pentru ca totuşi ciclul să se închidă, se stabilesc nişte noi indici δ, deduşi din d prin repartizarea uniformă a diferenţei $d_1 - d_1'$ între aceştia din urmă. Repartizarea se poate face fie prin progresie aritmetică, fie prin progresie geometrică, ultimul mod fiind cel mai uzitat. În acest scop, însemnând prin p factorul care face ca această diferenţă să apară de la o lună la alta, vom avea:

$$\delta_1 = d_1$$

$$\delta_2 = \frac{d_2}{p}$$

$$\delta_3 = \frac{d_3}{p^2}$$

.....

$$\delta_{12} = \frac{d_{12}}{p^{11}}$$

$$\delta'_1 = \frac{d'_1}{p^{12}} = \delta_1.$$

Ultima ecuaţie ne va permite determinarea lui p:

$$\log p = \frac{1}{12}(\log d'_1 - \log \delta'_1) = \frac{0{,}09212}{12} = 0{,}00768\ .$$

---

[1] *N. St. Georgescu,* ***Tehnica numerelor indice pentru nivelul general al preţurilor,*** *Bul. statistic al României, nr. 4, 1932; p. 116.*

Scanned with OKEN Scanner

TABELUL 146

Frecvenţele rapoartelor de la lună la lună

Valoarea raportului	Ian./ Dec.	Feb./ Ian.	Mar./ Feb.	Apr./ Mar.	Mai/ Apr.	Iun./ Mai	Iul./ Iun.	Aug./ Iul.	Sep./ Aug.	Oct./ Sep.	Nov./ Oct.	Dec./ Nov.
Peste 115	x	xxx	xxx		xxxx†	x	xxx		xxxx		x	xxxx
114-			xx		xxx O							
113-							x					
112-			x	xx							x	
111-			x O †					x				x
110-		x	x	x	x							x
109-		x										x O†
108-							x					
107-							x†					
106-		x O					O	x	†			
105-		x†			x		x	xx	x	x		
104-			xx					x				
103-					xx			x				x
102-	x	x				x	xx		x		x	x
101-			x	x	x							
100-						x		O	O	x	x	
99-	xx									x	†	
98-						x			x	xxx		x
97-		x		xx		x		†		o		
96-						x	x	xx	xx	x	xx	
95-	x			O †					x	xx		x
94-	x O			x			x	x		†	O	

Scanned with OKEN Scanner

Valoarea raportului	Ian./ Dec.	Feb./ Ian.	Mar./ Feb.	Apr./ Mar.	Mai/ Apr.	Iun./ Mai	Iul./ Iun.	Aug./ Iul.	Sep./ Aug.	Oct./ Sep.	Nov./ Oct.	Dec./ Nov.
93-						O					xx	
92-	†					†		x	x	x		
91-				x			x			x	x	x
90-	x			x		x					x	
89-				x				x				
88-	x											
87-			x									
86-												
85-						x					x	
Sub 85	xxx	xxx		xx		xxxx		xx		x	x	
Media	92,6	105,9	112,0	95,4	120,6	92,8	107,1	97,3	106,2	94,3	99,0	109,7
Mediana	94,4	106,2	111,7	95,6	113,7	93,7	106,4	100,0	100,4	97,5	94,7	109,9

Tabelul 145

Poziția medianei însemnată prin (O), *aceea a mediei aritmetice prin* (†).

Scanned with OKEN Scanner

Valorile succesive ale lui δ se obţin apoi prin ecuaţia:

$$\log \delta_i = \log d_i - (i-1)\log p.$$

După cum vedem, întrebuinţarea logaritmilor pentru toate operaţiile descrise mai sus se impune de la început. Amănuntele de calcul se văd în tabelul 147. În urmă, indicii δ se ajustează, reducându-i astfel ca media noilor indici să fie egală cu 100.

TABELUL 147

Determinarea indicilor de variaţie sezonală prin metoda lanţului de rapoarte de la lună la lună

Luna	Raportul median *c*	log *c*	log *d*	*i* log *p*	log δ_i	Indici ianuarie = 100	Indici ajustaţi
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ianuarie	94,4	1,97497	2,00000	0,	2,00000	100,0	91,4
Februarie	106,2	2,02612	2,02612	0,00768	2,01844	104,3	95,3
Martie	111,7	2,04805	2,07417	0,01535	2,05882	114,5	104,6
Aprilie	95,6	1,98046	2,05463	0,02303	2,03160	107,5	98,2
Mai	113,7	2,05576	2,11039	0,03071	2,07968	120,1	109,8
Iunie	93,7	1,97174	2,08213	0,03838	2,04375	110,6	101,1
Iulie	106,4	2,02694	2,10907	0,04606	2,06301	115,6	105,6
August	100,0	2,00000	2,10907	0,05374	2,05533	113,6	103,8
Septembrie	100,4	2,00173	2,11080	0,06141	2,04939	112,0	102,3
Octombrie	97,5	1,98900	2,09980	0,06909	2,03071	107,3	98,1
Noiembrie	94,7	1,97635	2,07615	0,07677	1,99938	99,9	91,3
Decembrie	109,9	2,04100	2,11715	0,08444	2,03271	107,8	98,5
Ianuarie	94,4	1,97497	2,09212	0,09212	2,00000	-	-
						1 313,2	**1 200,0**

Tabelul 146

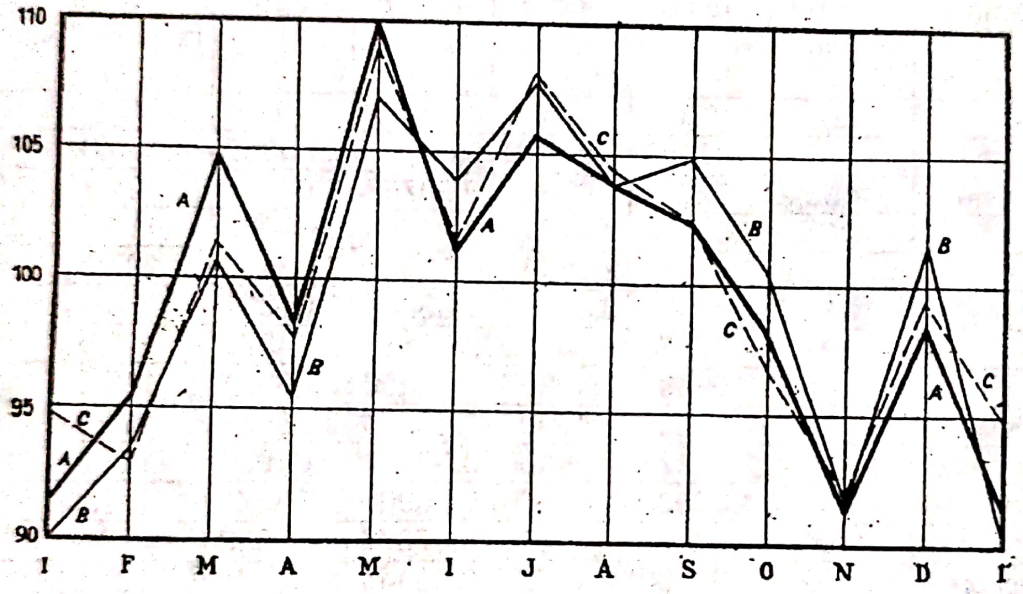
184. *Consideraţii asupra determinării variaţiilor sezonale.* Am aplicat dinadins procedeele de determinare a variaţiilor sezonale aceleiaşi serii statistice pentru ca astfel să punem în evidenţă faptul că, deşi dintre toate mişcările componente ale evoluţiei unui fenomen variaţia sezonală pare să se sprijine pe baze mai sigure, totuşi metode diferite nu conduc la rezultate identice (graficul 66). Cu toate acestea, nu putem să nu constatăm paralelismul care există între curbele reprezentând variaţia sezonală stabilită prin unul oarecare din cele trei procedee indicate.

Este un bun principiu ca, o dată obţinută curba de variaţie sezonală, să se caute şi o explicaţie a alurii sale în legătură cu factorii sezonali care influenţează fenomenul corespunzător. Această cercetare, în afară de faptul că

Scanned with OKEN Scanner

vine să completeze însăşi operaţia de determinare a variaţiilor sezonale, poate prezenta şi izolat un interes de studiu al acestor variaţii.

Dacă punem însă problema în acest din urmă mod, nu trebuie să uităm că diferenţele de la lună la lună pot fi pricinuite şi de inegalitatea numărului de zile ale diferitelor luni. Pentru a stabili, prin urmare, adevărata variaţie sezonală, este necesar să nu lucrăm cu datele globale, ci cu mediile zilnice. Şi aici este nevoie să ţinem seama nu de numărul zilelor dintr-o lună, ci de cel al zilelor *utile*. Dacă este vorba despre un fenomen meteorologic sau demografic, atunci evident că media zilnică va fi obţinută prin împărţirea datelor lunare la numărul de zile ale fiecărei luni. Pentru un fenomen economic, cum ar fi, de pildă, producţia industrială, datele lunare vor fi împărţite la numărul zilelor lucrătoare din luna respectivă. Apoi, mediile zilnice astfel obţinute vor fi supuse, pentru determinarea variaţiilor sezonale, uneia oarecare dintre metodele cunoscute.



A = Tabelul 147, B = Tabelul 144, C = Tabelul 141

Graficul 66. Variaţia sezonală a activităţii rafinăriilor de petrol din România.

Dacă însă variaţiile sezonale sunt determinate numai în scopul descompunerii mişcării fenomenului în componentele sale şi pentru acest lucru sunt utilizate totalurile lunare, atunci este nevoie ca şi variaţiile sezonale să fie determinate pornind de la acestea din urmă; diferenţele între numărul zilelor utile dintr-o lună vor fi considerate ca un factor de variaţie sezonală. Astfel se explică zigzagurile curbelor din graficul 66, care acuză în general un maximum pentru lunile de 31 de zile şi un minimum pentru celelalte.

185. *Componenta reziduală.* Procedeele expuse mai sus ne permit să determinăm pentru un anumit fenomen şi pentru o perioadă dată tendinţa de

Scanned with OKEN Scanner

lungă durată şi variaţiile sezonale. Ar urma, astfel, ca cea de-a treia necunoscută, mişcarea ciclică, să fie imediat calculată prin eliminarea celor două componente pe care le cunoaştem. Problema se pune global în acest mod. Vom vedea însă că intervin lucruri care o fac să nu mai apară sub această formă simplă.

Să admitem că am determinat tendinţa de lungă durată Y şi variaţiile sezonale v pentru o serie statistică în timp. Dacă fenomenul considerat n-ar fi supus acestor două componente, atunci mişcarea lui ar fi descrisă de produsul vY, adică vom avea: $y = vY$, unde v şi Y sunt valorile corespunzătoare lunii considerate. Cum însă acest lucru nu se întâmplă niciodată, din cauza mişcărilor care se suprapun celor două deja considerate, va exista un reziduu absolut $y - vY$ sau o componentă reziduală relativă $\frac{y - vY}{vY}$, care se socoteşte de obicei în procente:

$$r_1 = \frac{y \times 100}{vY} - 100. \qquad (150)$$

Dar acesta nu este singurul mod de a defini componenta reziduală relativă. Într-adevăr, putem adopta ca valoare normală valoarea tendinţei de lungă durată şi raporta reziduul absolut la Y. Obţinem:

$$r_2 = \frac{y \times 100}{Y} - v \ ^{1}. \qquad (150 \text{ bis})$$

Cum raportul dintre r_1 şi r_2 este egal cu v şi cum înseşi aceste valori nu sunt prea mari, dacă v nu variază prea mult de o parte şi de alta a valorii medii 100, ci între 90 şi 110, de pildă, atunci formulele precedente conduc la rezultate sensibil egale. De aceea, în acest din urmă caz, se adoptă în general formula (150 bis) care este mai comodă pentru calcul, deoarece nu necesită calcularea produsului vY.

În cazul datelor anuale, ambele formule devin:

$$r = \frac{y \times 100}{Y} - 100. \qquad (150 \text{ ter})$$

Aplicarea formulelor (150 bis) şi (150 ter) nu întâmpină nici o dificultate. Amănuntele de calcul se văd din tabelele 132, 133, 134, 138, 139.

186. *Componenta ciclică. Metoda lui Spencer.* Reprezentând grafic seria de valori a abaterilor reziduale, observăm că pe historiograma corespunzătoare nu mai putem distinge nici tendinţă de lungă durată, nici variaţii sezonale. Singura mişcare care mai poate fi desprinsă din grafic este mişcarea

---

[1] *v este socotit în unităţi simple în prima formulă şi în procente în ultima.*

Scanned with OKEN Scanner

ondulatorie. Nu trebuie însă crezut că aceasta se oglindeşte în componenta reziduală. Într-adevăr, mişcarea ondulatorie trebuie concepută ca un mecanism recurent şi în bună parte de alură regulată, prin urmare. Or, curba componentelor reziduale este departe de a avea această înfăţişare şi, dimpotrivă, prezintă oscilaţii din cele mai curioase (graficul 67). Aceste oscilaţii sunt efectul unei a patra componente a mişcării în timp a fenomenelor şi pe care am trecut-o sub tăcere atunci când am făcut analiza evoluţiei acestora, tocmai pentru a o prezenta în momentul în care ne putem da precis seama de existenţa ei. Este componenta întâmplătoare, rezultatul tuturor acţiunilor specifice fiecărui moment în parte şi care nu au nici o durată de un ordin apreciabil.

Dacă am ales termenul de componentă întâmplătoare pentru denumirea efectelor acestor cauze, nu înseamnă că ele înseşi sunt aleatorii; am putea cel mult spune că sunt sporadice. Termenul este motivat însă de aspectul în ansamblu al acestor abateri şi în care recunoaştem în parte caracterul variabilelor întâmplătoare.

Prin urmare, componenta reziduală cuprinde şi ea două componente: variaţia ciclică şi abaterea întâmplătoare. Pentru a face disjuncţia acestora, avem la îndemână un procedeu cunoscut. Toată problema constă din determinarea unei linii line care să urmărească pasul liniei agitate a componentei reziduale şi care să reprezinte variaţia ciclică. Pentru atingerea acestui scop putem întrebuinţa exact aceleaşi metode ca şi în cazul determinării tendinţei de lungă durată, adică: trasarea curbei cu mâna liberă, metoda mediilor mobile sau metoda interpolării.

Trebuie să recunoaştem că ultimul procedeu este destul de complicat şi uneori imposibil de aplicat din cauza neregularităţii variaţiilor ondulatorii[1]. Metoda mediilor mobile este cea mai preferată. Ea se impune cu atât mai mult în cazul special al determinării variaţiilor sezonale, cu cât componenta întâmplătoare, care trebuie eliminată, se prezintă sub aspectul unor adevărate abateri ce se manifestă când într-un sens, când într-altul. Numărul de termeni asupra cărora trebuie să poarte media mobilă trebuie să fie cât mai mic cu putinţă pentru a nu nivela prea mult variaţiile ciclice (§171). O medie mobilă pe 5 sau 7 termeni conduce la un rezultat mulţumitor.

Nu trebuie însă neglijat faptul că metoda mediilor mobile tinde să transforme orice mişcare, fie ea chiar pozitiv curbilinie, într-o alta, de formă mai puţin accidentată, însă mult mai aproape de o linie dreaptă. Or, cum mişcarea ciclică va fi reprezentată de o serie de unde, prin aplicarea metodei mediilor mobile vom micşora *amplitudinea* acestor ondulaţii, adică vom coborî maximele şi vom ridica minimele. Pentru acest motiv, s-a căutat să se imagineze alte procedee asemănătoare metodei mediilor mobile, care însă să nu mai prezinte neajunsul semnalat. Aceste procedee sunt din cele mai variate şi numărul mare al lor ne împiedică să le pomenim aici pe toate. Ele se numesc, cu un termen

[1] *N. St. Georgescu,* ***Le problème de la recherche des composantes cycliques d'un phenomène.***

Scanned with OKEN Scanner

general, metode de netezire[1] şi nu prezintă decât deosebiri de amănunt una faţă de alta.

TABELUL 148

Cacului componentei ciclice prin metoda lui Spencer pentru 21 de termeni[2]

LUNA	Componenta reziduală	Suma ponderată s_0	Suma mobilă (7 termeni)	Suma mobilă (5 termeni)	Suma mobilă (5 termeni)	Componenta ciclică	Componenta întâmplătoare
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1925							
Ianuarie	-5,7	.	.	.	.	.	.
Februarie	-1,7	.	.	.	.	.	.
Martie	+0,6	.	.	.	.	.	.
Aprilie	-5,2	-1,4	.	.	.	.	.
Mai	-5,6	+3,0	.	.	.	.	.
Iunie	+2,9	-10,4	.	.	.	.	.
Iulie	-8,3	-19,5	-77,9	.	.	.	.
August	-14,8	-44,1	-38,2	.	.	.	.
Septembrie	+1,7	-18,5	-37,7	-177,7	.	.	.
Octombrie	-3,8	+13,0	-24,0	-45,7	.	.	.
Noiembrie	+13,5	+38,3	+0,1	+64,4	+222,7	+0,64	+12,9
Decembrie	+0,4	+3,5	+54,1	+157,1	+716,4	+2,05	-1,7
1926							
Ianuarie	+2,9	+3,3	+71,9	+224,6	+1.181,1	+3,38	-0,5
Februarie	+0,1	+4,6	+55,0	+316,0	+1.660,3	+4,74	-4,6
Martie	+12,0	+9,9	+43,5	+419,0	+2.232,5	+6,38	+5,6
Aprilie	+6,8	-0,7	+91,5	+543,6	+3.013,1	+8,61	-1,8
Mai	-3,0	-3,9	+157,1	+729,3	+4.018,8	+11,48	-14,5
Iunie	+20,6	+26,8	+196,5	+1.005,2	+5.199,8	+14,86	+5,7
Iulie	+20,4	+51,5	+240,7	+1.321,7	+6.641,7	+18,98	+1,4
August	+25,2	+68,9	+319,4	+1.600,0	+7.554,8	+21,58	+3,6
Septembrie	+19,8	+44,0	+408,0	+1.785,5	+8.303,5	+23,72	-3,9
Octombrie	+28,3	+54,1	+435,4	+1.842,4	+8.516,9	+24,33	+4,0
Noiembrie	+24,7	+78,0	+382,0	+1.753,9	+8.146,5	+23,28	+1,4
Decembrie	+28,5	+84,7	+297,6	+1.535,1	+7.290,2	+20,83	+7,7

[1] Graduation, *în engleză.*

[2] *Tot pentru motivul lungimii sale, nu publicăm decât o parte a tabelului la care dă naştere aplicarea metodei la întreaga serie din tabelul 136. Pentru exemplificare numai, acesta este însă suficient.*

Scanned with OKEN Scanner

(continuare tab. 148)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1927							
Ianuarie	+26,6	+54,2	+230,9	+1.229,6	+6.130,5	+17,52	+9,1
Februarie	+3,0	-1,9	+189,2	+929,2	+4.919,9	+14,06	-11,6
Martie	+3,8	-15,5	+129,9	+682,7	.	.	.
Aprilie	+2,2	-22,7	+81,6	+543,3	.	.	.
Mai	+13,6	+12,4	+51,1	.	.	.	.
Iunie	-0,2	+18,7	+91,5	.	.	.	.
Iulie	+17,9	+36,4	.	.	.	.	.
August	+13,8	+23,7	.	.	.	.	.
Septembrie	+8,6	+38,5	.	.	.	.	.
Octombrie	+10,8	.	.	.	.	.	.
Noiembrie	+16,8	.	.	.	.	.	.
Decembrie	+3,5	.	.	.	.	.	.

Tabelul 139

Vom menționa pe cel mai răspândit și care este cunoscut sub numele de procedeul de 21 de termeni al lui J. SPENCER[1]. Pentru a determina componenta ciclică prin acest procedeu, se fac următoarele operații asupra seriei componentelor reziduale:

1. Se calculează o sumă mobilă ponderată după formula:

$$s_0 = -u_{-3} + u_{-1} + 2u_0 + u_1 - u_3 \qquad (151)$$

unde $...u_{-3},u_{-2},u_{-1},u_0,u_1,u_2,u_3...$ sunt termenii consecutivi ai seriei componentei reziduale. Suma s_0 se centrează în dreptul termenului u_0 (tabelul 148, coloana 3). Astfel, termenului crud - 5,2 îi va corespunde suma: + 5,7 + 0,6 - 2 x 5,2 - 5,6 + 8,3 = - 1,4.

2. Cu ajutorul acestor rezultate se calculează o sumă mobilă pe 7 termeni (coloana 4).

3. Se ia apoi o sumă mobilă pe cinci termeni (coloana 5) și, din rezultat, o nouă sumă mobilă pe cinci termeni (coloana 6).

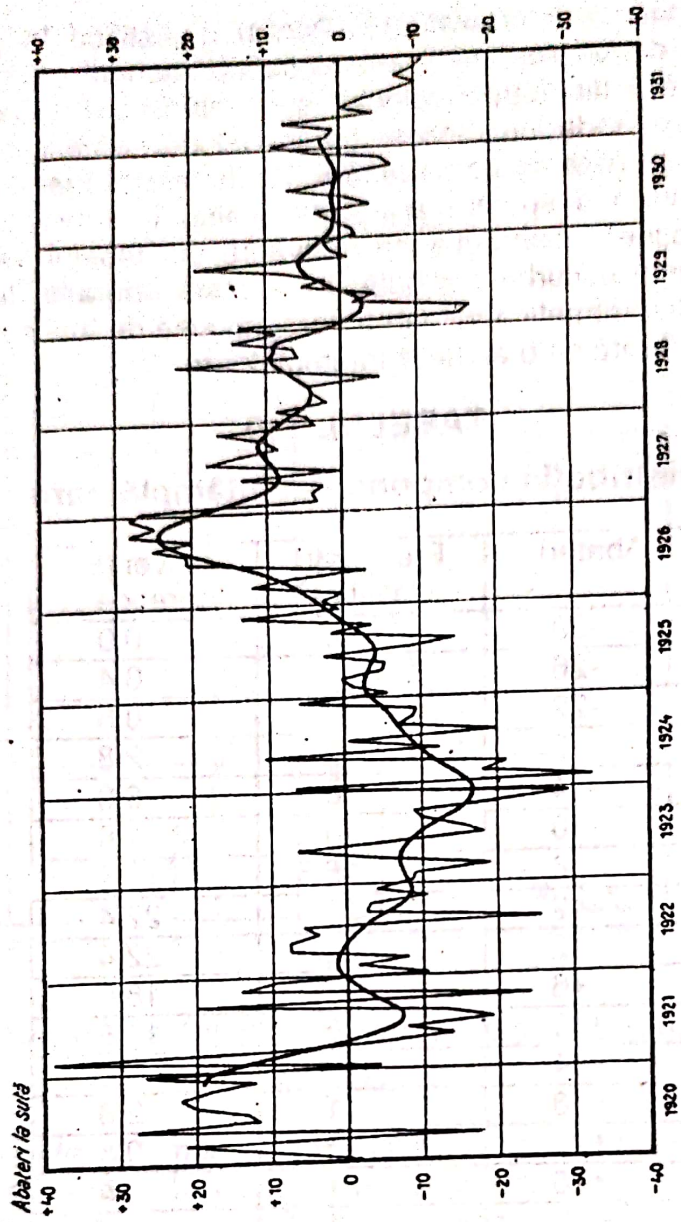
4. Rezultatul se împarte cu 350 (coloana 7).

Valorile astfel obținute constituie o excelentă evaluare a componentelor ciclice. Reprezentarea grafică a seriei obținute ne pune în evidență mișcarea ciclică a fenomenului considerat care, așa cum am mai spus, se prezintă sub forma unei linii ondulate line (graficul 67).

---

[1] *J. Spencer, **On the Graduation of the Rates of Sickness and Mortality**, Journal of the Institute of Actuaries, vol. XXXVIII, p. 334.*

Scanned with OKEN Scanner



Tabelul 148

Graficul 67. Mișcarea ciclică a activității rafinăriilor de petrol din România, calculată prin metoda lui *Spencer pentru 21 de termeni.*

Celelalte metode de netezire a unei serii statistice constau dintr-o suprapunere analoagă de sume mobile simple sau ponderate. Ele pot fi aplicate nu numai componentelor reziduale, ci și seriei efective. În acest din urmă caz, dacă se are grijă ca printre sumele mobile să figureze și una de 12 termeni, prin care să se elimine influența abaterilor întâmplătoare, rezultatul va oglindi mișcarea ciclică suprapusă celei de lungă durată[1].

[1] *De fapt, toate aceste procedee se pot aplica pentru netezirea oricărei serii statistice; o extinsă întrebuințare a lor se face în legătură cu tablele de mortalitate, chestiune care a fost chiar originea primelor metode de netezire imaginate.*

Scanned with OKEN Scanner

187. *Componenta întâmplătoare*. Pentru a justifica în parte termenul întrebuințat pentru denumirea celei de-a patra componente a mișcării în timp, vom face o analiză a distribuției valorilor acesteia. În cazul seriei din tabelul 136, întrebuințând metoda lui Spencer pentru *21 de termeni*, obținem 124 de abateri întâmplătoare care se calculează în modul arătat mai sus (tabelul 148, coloana 8). Făcând o despuiere a acestor valori în raport cu o scară de intervale egale, obținem distribuția de frecvențe din tabelul 149. Pe această distribuție vom ajusta o curbă normală; în această operație de ajustare vom adopta pentru media aritmetică valoarea zero, impusă de însuși faptul că privim componenta considerată ca o abatere întâmplătoare[1].

TABELUL 149

Distribuția componentei întâmplătoare

Abateri	Frecvențe statistice	Frecvențe teoretice
-30	.	0,0
-26	.	0,4
-22	2	0,8
-18	2	2,8
-14	6	6,0
-10	10	11,6
-6	15,5	18,0
-2	32,5	22,4
+2	22	22,4
+6	20	18,0
+10	5	11,6
+14	4	6,0
+18	1	2,8
+22	2	0,8
+26	1	0,4
+30	1	0,0
	124	**124,0**

Tabelul 148

Ecuația curbei ajustate este:

$$y = \frac{124}{8{,}56\sqrt{2\pi}} e^{-\frac{x^2}{2\times 8{,}56^2}},$$

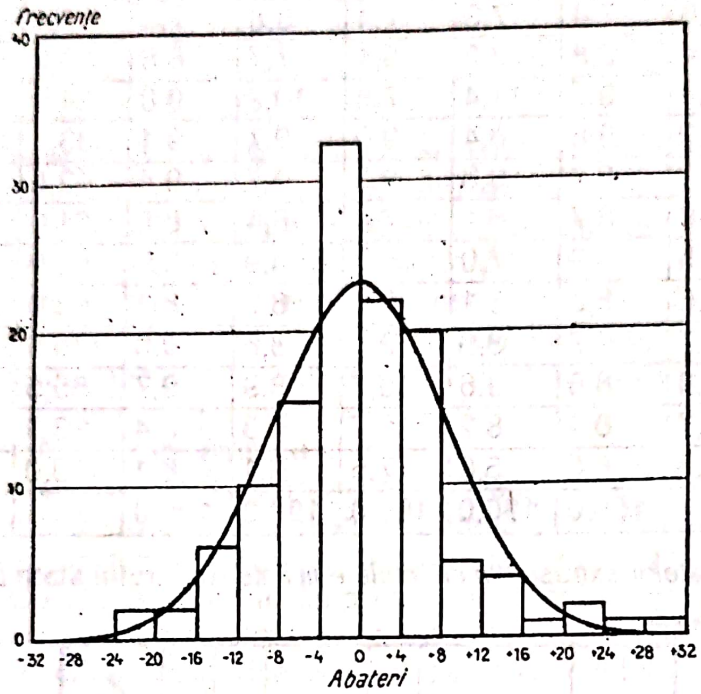
iar comparația grafică între frecvențele teoretice și cele statistice este făcută pe graficul 68. Lipsa unei concordanțe categorice între acestea (P' = 0,076971)

[1] *Valoarea mediei aritmetice statistice este - 0,14 și cum, pentru aceeași distribuție σ = 8,56, suntem evident în limitele variației statistice a unei medii teoretice nule.*

Scanned with OKEN Scanner

trebuie pusă în bună parte şi pe seama faptului că eşantionul nu conţine un număr suficient de observaţii. Analize mai amănunţite au condus totuşi la evidenţierea unei slabe asimetrii a distribuţiilor componentelor întâmplătoare.

Aspectul destul de apropiat al acestor distribuţii faţă de curba normală constituie una din cele mai întemeiate justificări ale definiţiei hazardului, aşa cum a fost formulată de H. POINCARÉ (§ 91).



Tabelul 149

Graficul 68. Comparaţie între curba normală şi distribuţia componentei întâmplătoare.

188. *O aplicaţie a formei cumulative a seriilor în timp*. Revenind asupra formei cumulative a seriilor în timp, vom expune una din principalele probleme pentru care această formă constituie un mijloc de interpretare foarte des întrebuinţat. Este vorba despre comparaţia dintre evoluţia proiectată a unui fenomen şi realizările în aceeaşi direcţie (§ 13).

Să considerăm încasările vamale lunare exprimate în procente faţă de totalul anual al încasărilor (tabelul 150). Cu ajutorul acestor date vom calcula în primul rând un indice lunar mijlociu al încasărilor. Seria sub formă cumulativă a acestui indice constituie seria proiect[1]. Diferenţele dintre termenii acesteia din urmă şi încasările din 1930 (exprimate de data aceasta în procente faţă de valoarea prevăzută în bugetul anului respectiv) ne conduc la valorile deficitelor lunare efective.

---

[1] *Avem de-a face, de fapt, cu un calcul de variaţie sezonală. Valoarea indicelui neajustat este egală cu mediana procentelor lunii respective; ajustarea se face proporţional, pentru a obţine un total egal cu 100.*

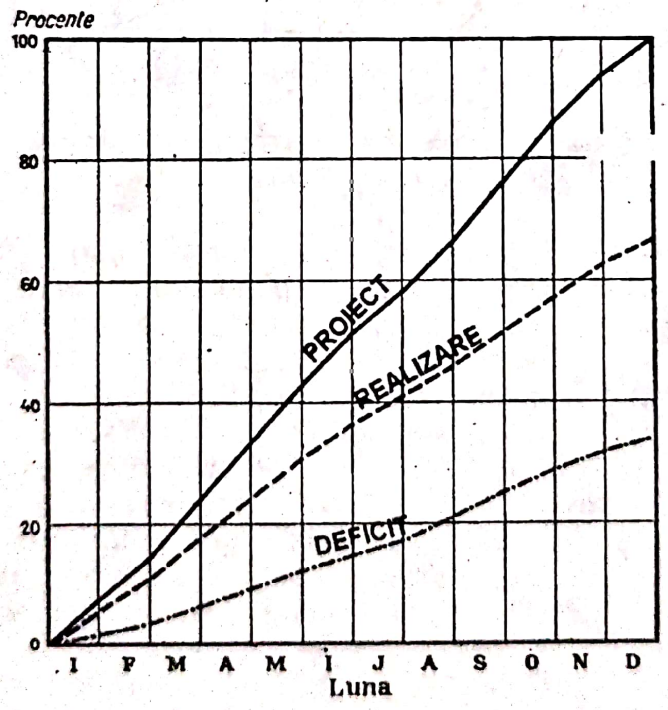
Scanned with OKEN Scanner

TABELUL 150

Încasările din taxele vamale în România (1926 - 1930)

Luna	Procente față de total					Indice mijlociu	Indice cumulativ	Realizări (1930)	Deficit (1930)
	1926	1927	1928	1929	1930				
Ianuarie	6,6	7,4	7,3	7,4	8,8	7,4	7,4	5,8	1,6
Februarie	6,4	6,8	8,5	5,4	7,7	6,8	14,2	10,9	3,3
Martie	7,9	9,7	11,4	7,8	10,2	9,8	24,0	17,7	6,3
Aprilie	6,0	9,8	8,4	9,0	9,7	9,1	33,1	24,1	9,0
Mai	8,1	8,9	9,3	9,3	9,7	9,4	42,5	30,6	11,9
Iunie	8,9	8,7	8,1	7,4	8,4	8,5	51,0	36,2	14,8
Iulie	7,5	6,9	7,0	6,7	6,9	6,9	57,9	40,8	17,1
August	8,6	8,9	8,1	8,3	6,7	8,1	66,0	45,2	20,8
Septembrie	10,7	9,7	8,5	9,8	8,6	9,8	75,8	50,9	24,9
Octombrie	10,3	8,8	9,6	10,7	8,9	9,7	85,5	56,8	28,7
Noiembrie	10,7	8,4	8,2	10,7	8,3	8,4	93,9	62,3	31,6
Decembrie	8,3	6,0	5,6	7,5	6,1	6,1	100,0	66,3	33,7
	100,0	**100,0**	**100,0**	**100,0**	**100,0**	**100,0**			

Calculat pe baza datelor extrase din diferitele ediții ale **Anuarului statistic al României**.



Tabelul 150

Graficul 69. Comparație între mersul încasărilor din taxele vamale și evaluările bugetare.

Capitolul XIV

Covariația

189. *Generalități.* La fel ca și în cazul distribuțiilor de frecvență, comparația seriilor în timp a două sau mai multe fenomene constituie partea cea mai importantă a analizei la care acestea pot fi supuse. Seriile în timp sunt comparate între ele în scopul de a stabili, în primul rând, o concordanță între variațiile prezentate de fenomenele corespunzătoare. Concordanța dintre evoluțiile a două fenomene nu trebuie concepută ca o legătură rigidă nici chiar în ceea ce privește sensul variațiilor acestora. Privind chestiunea din punctul de vedere formal, problema concordanței variațiilor în timp se prezintă sub același aspect ca și corelația dintre atribute.

Cu toate acestea, ambele probleme se deosebesc fundamental între ele, consecință a faptului accentuat suficient în prima parte a capitolului precedent: evoluțiile în timp ale fenomenelor nu pot fi tratate pe baza rezultatelor teoriei eșantioanelor. Ceva mai mult, în timp ce concordanța variațiilor sau prezența concomitentă a două atribute ne conduc întotdeauna la stabilirea logică a unei legi stocastice - chiar dacă aceasta nu poate fi complet explicată - acest procedeu nu poate fi legitim în cazul evoluțiilor în timp. Într-adevăr, teoretic vorbind, orice evoluție în timp este o funcție de acesta din urmă. Eliminând, prin urmare, variabila timpului între două din aceste funcții, obținem o legătură *funcțională* între fenomenele respective, oricare ar fi acestea, chiar dacă existența unei cauzalități între manifestările lor ar constitui o imposibilitate demonstrată. Așadar, legătura funcțională prin intermediul timpului se poate stabili între fenomenele cele mai străine unul de altul; cauzalitatea nu trebuie însă identificată cu această legătură formală.

Pe de altă parte, chiar dacă două fenomene s-ar găsi, prin firea lucrurilor, sub influența aceluiași ansamblu de cauze, efectul acestora din urmă poate - și chiar e logic să fie așa - să nu fie același de la o epocă la alta. Intervenția voită sau inconștientă a oamenilor schimbă cu timpul nu numai evoluția fenomenelor, dar chiar sistemul de cauzalitate dintre ele. Ne gândim în special la fenomenele sociale pentru care seriile în timp constituie aproape singurul mijloc de analiză. Chiar pentru fenomenele asupra cărora omul poate acționa mai puțin sau deloc - dacă mijlocul de stabilire a unei legături cauzale este numai comparația evoluțiilor în timp respective - concordanța sau lipsa acesteia nu pot fi privite ca un aspect al realității. Izolarea cauzelor care influențează fenomenele din această categorie constituie în majoritatea cazurilor o imposibilitate; altfel nici n-am recurge la studiul lor prin intermediul seriilor în timp.

Scanned with OKEN Scanner

Analiza legăturii stocastice dintre două fenomene făcută cu ajutorul seriilor în timp este de fapt un refugiu al spiritului omenesc, refugiu la care suntem obligaţi să recurgem din cauza imposibilităţii materiale a unei experimentări şi mai comode şi mai ştiinţifice în câmpul anumitor fenomene. Se mai întâmplă, din nenorocire, ca tocmai fenomenele ce interesează în gradul cel mai înalt ştiinţele sociale să nu se preteze disecţiilor care ne-ar scuti de a le studia prin seriile în timp.

Ne dăm acum seama de ce problema legăturii dintre seriile în timp se pune în fond cu totul altfel decât aceea a corelaţiei şi de ce spiritul critic al lui LUCIEN MARCH l-a condus pe acesta să propună pentru această nouă problemă a statisticii matematice termenul de *covariaţie*, în opoziţie cu şcoala anglo-saxonă care întrebuinţează tot termenul de corelaţie. Şi deşi această din urmă şcoală nu confundă cele două noţiuni de corelaţie şi de covariaţie, întrebuinţarea aceleiaşi terminologii şi pentru un caz şi pentru altul a făcut ca foarte mulţi cercetători să ajungă la o greşită interpretare a rezultatelor numerice obţinute prin aplicarea procedeelor descrise în paragrafele următoare.

190. *Indici şi coeficienţi de covariaţie diferenţială*. Existenţa covariaţiei dintre două fenomene se poate stabili pe cale grafică chiar. Nu ne mai oprim asupra acestui procedeu, ci vom trece la descrierea indicilor numerici, menţionând însă că el trebuie să preceadă în practică analiza prin indici, pentru a evita astfel calcule destul de lungi în unele cazuri şi care nu ne-ar conduce decât la confirmarea impresiei făcută de grafic.

Pentru ilustrarea procedeelor de calcul care ne conduc la diferiţi indici de covariaţie, ne vom referi la seriile din tabelul 151.

TABELUL 151

Covariaţia diferenţială dintre indicele preţurilor de gros şi producţia de fontă[1] (Statele Unite, 1900 - 1915)

Anii	Indicele preţurilor de gros x	Producţia de fontă (mil. tone) y	Diferenţe		Produse		Pătrate	
			Δx	Δy	$\Delta x \Delta y$		$(\Delta x)^2$	$(\Delta y)^2$
(1)	(2)	(3)	(4)		(5)		(6)	
1900	91	14,0	.	.	.	.	.	.
1901	89	16,1	-2	+2,1	.	-4,2	4	4,41
1902	94	18,1	+5	+2,0	+10,0	.	25	4,00
1903	96	18,3	+2	+0,2	+0,4	.	4	0,04
1904	96	16,8	0	-1,5	.	.	.	2,25
1905	96	23,4	0	+6,6	.	.	.	43,56

[1] *Indicii de preţ sunt cei publicaţi de Bureau of Labor Statistics.*

Scanned with OKEN Scanner

(continuare tab. 151)

(1)	(2)	(3)	(4)		(5)		(6)	
1906	100	25,7	+4	+2,3	+9,2	.	16	5,29
1907	105	26,2	+5	+0,5	+2,5	.	25	0,25
1908	101	16,2	-4	-10,0	+40,0	.	16	100,00
1909	109	26,2	+8	+10,0	+80,0	.	64	100,00
1910	113	27,7	+4	+1,5	+6,0	.	16	2,25
1911	104	24,0	-9	-3,7	+33,3	.	81	13,69
1912	111	30,2	+7	+6,2	+43,4	.	49	38,44
1913	112	31,5	+1	+1,3	+1,3	.	1	1,69
1914	110	23,7	-2	-7,8	+15,6	.	4	60,84
1915	113	30,4	+3	+6,7	+20,1	.	9	44,89
	1640	**368,5**			**+260,8**	**-4,2**	**314**	**421,60**

Datele extrase din ***Annuaire Statistique de la France****, 1930, p. 295 și 380.*

Trebuie să observăm de la început că, pentru a obține o măsură a covariației, putem să luăm ca bază fie concordanțele - eventual discordanțele - diferențelor între termenii consecutivi ai celor două serii, fie concordanțele între abaterile acestora, calculate în raport cu valorile medii corespunzătoare. Studiem în primul caz *covariația diferențială*, în al doilea, *covariația tangențială*. Diferența dintre aceste două feluri de covariații a fost pusă în evidență tot de LUCIEN MARCH. Ele corespund la două noțiuni deosebite. Covariația diferențială constă din concordanța variațiilor fenomenelor de la o epocă la alta, în timp ce covariația tangențială traduce concordanța între direcțiile globale ale celor două fenomene. Este clar acum că se pot imagina cazuri în care cele două feluri de covariație să fie complet diferite; March a dat chiar exemple în care una din covariații este nulă, iar cealaltă acuză un grad foarte pronunțat[1].

Ne vom ocupa în primul rând de covariația diferențială. În conformitate cu definiția acesteia, vom calcula diferențele succesive ale termenilor seriilor considerate (coloana 4). Ori de câte ori două din aceste diferențe vor fi de același semn, vom avea o concordanță de variație în intervalul respectiv. Un indice foarte simplu va fi obținut dacă din numărul concordanțelor c, vom scădea pe acel al discordanțelor d, iar în urmă vom raporta rezultatul la numărul diferențelor:

$$i = \frac{c - d}{n}, \tag{152}$$

Acest coeficient a fost introdus de FECHNER[2]. Azi este cunoscut sub numele de *indicele simplu de covariație diferențială*. Valoarea lui poate să

[1] *L. March,* ***Comparaison numérique des courbes statistiques****, Journal de la Société de Statistique de Paris, aug. 1905, p. 270.*

[2] *G. Fechner,* ***Kollektivmasslehre****, Leipzig, 1897.*

Scanned with OKEN Scanner

varieze de la -1 la +1, limitele fiind atinse în cazul când toate diferenţele dau naştere fie numai la concordanţe, fie numai la discordanţe.

Indicele *i* este nul în cazul când numărul concordanţelor este egal cu al discordanţelor şi când ar fi hazardat să concludem la existenţa unei covariaţii: covariaţia este în acest din urmă caz nulă.

Totuşi, indicele *i* nu constituie o măsură prea fericită a covariaţiei, deoarece acesta este bazat pe principiul că orice concordanţă este anulată de o discordanţă. Or, este clar că şi în ceea ce priveşte concordanţele şi discordanţele variaţiilor există o scară graduală şi că, prin urmare, nu toate acestea pot fi privite ca identice.

Se remediază acest defect dacă individualizăm *gradul* unei concordanţe sau discordanţe prin produsul $\Delta x \cdot \Delta y$. Indicele care decurge poartă numele de *indicele ponderat de covariaţie diferenţială* şi este dat de formula:

$$I = \frac{C - D}{C + D}, \tag{153}$$

unde C reprezintă suma produselor $\Delta x \cdot \Delta y$ pozitive, iar D valoarea absolută a sumei produselor negative.

Referindu-ne la exemplul din tabelul 151, vom avea:

$$i = \frac{12 - 1}{15} = 0{,}73 \quad [1] \qquad I = \frac{260{,}8 - 4{,}2}{260{,}8 + 4{,}2} = 0{,}97.$$

Diferenţa dintre aceste două valori provine tocmai din faptul că singura discordanţă care există influenţează mai mult indicele simplu decât pe cel ponderat.

Ambii aceşti indici prezintă însă un mare dezavantaj; ei pot atinge limitele ± 1 chiar atunci când covariaţia nu este ideală, adică chiar când diferenţele Δx şi Δy nu sunt matematic proporţionale.

Astfel, dacă în exemplul de mai sus ne referim numai la intervalul 1905-1915, vom avea:

$$i = \frac{10}{10} = +1 \qquad I = \frac{250{,}4}{250{,}4} = +1$$

şi totuşi nu suntem în cazul unei legături rigide. Acest defect al indicilor de covariaţie nu se mai întâlneşte în *coeficientul de covariaţie diferenţială*. Coeficientul de covariaţie este construit după structura coeficientului pearsonian de corelaţie:

---

[1] *În cazul când una din diferenţele Δ este nulă, intervalul respectiv nu este considerat nici ca o concordanţă, nici ca o discordanţă, ci este lăsat deoparte în calculul lui* i.

Scanned with OKEN Scanner

$$k = \frac{\sum \Delta x . \Delta y}{\sqrt{\sum (\Delta x)^2 \sum (\Delta y)^2}} \qquad (154)$$

și nu poate fi egal cu ±1 decât atunci când diferențele Δy ar varia direct proporțional cu Δx. Calculul lui k pentru tabelul 151 nu întâmpină nici o dificultate:

$$k = \frac{260,8 - 4,2}{\sqrt{314 \times 421,60}} = \frac{256,6}{363,8} = 0,71 .$$

191. *Indici și coeficienți de covariație tangențială.* Pentru a măsura covariația tangențială, vom utiliza abaterile termenilor față de media lor aritmetică. În cazul seriilor considerate, avem:

$$M_x = 102,5 \quad M_y = 23,03.$$

Abaterile $x - M_x$, $y - M_y$ sunt înscrise în coloana (2) din tabelul 152. Coeficienții de covariație tangențială sunt calculați pe aceeași cale ca și cei de covariație diferențială, utilizând însă abaterile în raport cu media aritmetică, în locul diferențelor $\Delta x, \Delta y$.

TABELUL 152

Covariația tangențială dintre indicele prețurilor de gros și producția de fontă

Ani	Abateri		Produse		Pătrate	
	$x-M_x$	$y-M_y$	$(x-M_x)(y-M_y)$		$(x-M_x)^2$	$(y-M_y)^2$
(1)	(2)		(3)		(4)	
1900	-11,5	-9,03	+103,845	.	132,25	81,54
1901	-13,5	-6,93	+93,555	.	182,25	48,02
1902	-8,5	-4,93	+41,905	.	72,25	24,30
1903	-6,5	-4,73	+30,745	.	42,25	22,37
1904	-6,5	-6,23	+40,495	.	42,25	38,81
1905	-6,5	+0,37	.	-2,405	42,25	0,14
1906	-2,5	+2,67	.	-6,675	6,25	7,13
1907	+2,5	+3,17	+7,925	.	6,25	10,05
1908	-1,5	-6,83	+10,245	.	2,25	46,65
1909	+6,5	+3,17	+20,605	.	42,25	10,05
1910	+10,5	+4,67	+49,035	.	110,25	21,81
1911	+1,5	+0,97	+1,455	.	2,25	0,94
1912	+8,5	+7,17	+60,945	.	72,25	51,41
1913	+9,5	+8,47	+80,465	.	90,25	71,74

Scanned with OKEN Scanner

(continuare tab. 152)

(1)	(2)		(3)		(4)	
1914	+7,5	+0,67	+5,025	.	56,25	0,45
1915	+10,5	+7,37	+77,385	.	110,25	54,32
			+623,630	**-9,080**	**1 012,00**	**489,73**

Tabelul 151

Indicele simplu de covariaţie tangenţială, *j*, se calculează pe baza formulei (152), concordanţele şi discordanţele fiind stabilite în raport cu abaterile $x\text{-}M_x$, $y\text{-}M_y$. Indicele ponderat de covariaţie tangenţială, *J*, are aceeaşi structură ca *I* şi se obţine pe baza formulei (153). În cazul tabelului 152, avem:

$$j = \frac{14-2}{16} = 0{,}75 \qquad J = \frac{623{,}63-9{,}08}{623{,}63+9{,}08} = 0{,}97.$$

În fine, coeficientul de covariaţie tangenţială este dat de formula:

$$c = \frac{\sum(x-M_x)(y-M_y)}{\sqrt{\sum(x-M_x)^2 \cdot \sum(y-M_y)^2}} \qquad (155)$$

care-l face să se confunde formal cu coeficientul de corelaţie *r*. În cazul tabelului 152:

$$c = \frac{623{,}63-9{,}08}{\sqrt{1012 \times 489{,}73}} = \frac{614{,}55}{703{,}99} = 0{,}87.$$

Asupra indicilor şi coeficientului de covariaţie tangenţială avem de făcut aceleaşi observaţii pe care le-am făcut în cazul covariaţiei diferenţiale, pentru care motiv nici nu le mai repetăm.

Totuşi, nici *c*, nici *k* nu constituie nişte buni indici ai covariaţiei decât în cazul când între diferenţele Δ sau abaterile în raport cu media există o legătură de formă oarecum liniară. Pentru cazul când această din urmă condiţie nu este îndeplinită - caz ce se prezintă destul de rar în practică - s-au construit indici de o structură diferită. DI ŞERBAN GHEORGHIU a indicat un procedeu general care să permită stabilirea unui indice de covariaţie, oricare ar fi structura algebrică a legăturii dintre seriile considerate[1].

Foarte mulţi autori, printr-o confuzie pe care am menţionat-o, vorbesc de o *abatere tip* a coeficienţilor de covariaţie pe care o identifică grăbit cu aceea a coeficientului de corelaţie (§144), calculând, de pildă, pentru coeficientul *c* o abatere tip după formula:

$$\frac{1-c^2}{\sqrt{n}}.$$

[1] *Şerban Gheorghiu, **Essai sur les finances d'après-guerre de la Roumanie** (Thèse), Journal de la Société de Statistique de Paris, 1926, p. 309 şi 355.*

Scanned with OKEN Scanner

Absurditatea acestei identificări este evidentă. Coeficientul *r* are o valoare teoretică pe care noi o evaluăm statistic prin intermediul unui eşantion al colectivităţii considerate. În jurul valorii teoretice, valorile statistice se grupează, sub anumite condiţii, după o curbă normală pentru care abaterea tip are semnificaţia precisă pe care o cunoaştem. Unde este eşantionul în cazul seriilor în timp şi ce valoare teoretică este evaluată prin oricare din coeficienţii de covariaţie?

Singur numitorul expresiei de mai sus poate fi în parte justificat, dacă în loc de a o privi ca o abatere tip o vom considera ca un *coeficient de precizie*: cu cât vom avea la dispoziţie mai multe date, cu atât valorile coeficienţilor de covariaţie deduşi vor fi mai puţin discutabile.

În ceea ce priveşte semnificaţia valorilor acestor coeficienţi, trebuie să spunem de la început că nu putem conchide la existenţa unei covariaţii nete în cazul când valoarea absolută a lor este mai mică decât 0,3.

192. *Componentele evoluţiei în timp şi problema covariaţiei.* Covariaţia a două fenomene nu apare întotdeauna chiar din examenul seriilor efective, aşa cum s-a întâmplat în exemplul examinat. De fapt, covariaţia stabilită astfel nici nu prezintă un prea mare interes.

Covariaţia tangenţială măsurată pe baza datelor brute poate să acuze un grad de intensitate mai ridicat decât cel real: chiar dacă fenomenele nu stau într-un raport cauzal, atâta vreme cât au o tendinţă evolutivă liniară, covariaţia tangenţială a lor va fi supraestimată de oricare din indicii de măsură a acesteia. (Aşa se şi explică de ce, pentru seriile din tabelul 151, coeficientul de covariaţie tangenţială e mult superior celui de covariaţie diferenţială.)

Alteori, fenomene a căror dependenţă nu mai poate fi discutată nu conduc la valori semnificative pentru coeficienţii de covariaţie pentru motivul că raportul lor de cauzalitate se manifestă numai asupra uneia din multiplele componente ale mişcării lor în timp. De aici, nevoia de a cerceta covariaţia dintre mişcările fenomenelor descompuse în componentele lor. Şi nu greşim afirmând că mai întotdeauna covariaţia mişcărilor ondulatorii constituie o probă corectă a dependenţei dintre două fenomene.

Exemplul din tabelul 153 ne va permite să verificăm afirmaţiile de mai sus. Este vorba despre covariaţia dintre nupţialitate şi disponibilităţile materiale ale populaţiei. O legătură cauzală între aceste două variabile se concepe firesc *a priori*.

Vom adopta importul în valoare ca indice al disponibilităţilor materiale ale populaţiei şi vom examina problema anunţată, servindu-ne de datele referitoare la România pentru perioada 1888 - 1914. Din însăşi reprezentarea grafică a seriilor se vede clar că evoluţia lor în timp nu se face într-un acord prea pronunţat (graficul 70). Pentru datele brute, obţinem:

$$k = 0{,}41 \qquad c = 0{,}86\ ^{1}.$$

[1] *Valoare care supraestimează, după cum ştim, gradul real de covariaţie.*

Scanned with OKEN Scanner

Căutând să studiem covariația dintre mișcările ondulatorii ale fenomenelor considerate, vom calcula componentele ondulatorii corespunzătoare, utilizând pentru determinarea tendinței de lungă durată metoda mediilor mobile pe 5 ani (§171). Componentele r_x, r_y reziduale sunt înscrise în coloana 4 (tabelul 153).

TABELUL 153

Covariația dintre nupțialitate și import

(România, 1888 - 1914)

Anii	Căsătorii (mii) x	Import (mil. lei) y	Componente reziduale		Produse		Pătrate		Abateri reduse	
			r_x	r_y	$r_x r_y$		r_x^2	r_y^2	r_x / σ_1	r_y / σ_2
(1)	(2)	(3)	(4)		(5)		(6)		(7)	
1888	38,3	310,4	.	.	.	.	.	.	.	.
1889	41,1	367,9	.	.	.	.	.	.	.	.
1890	38,6	362,8	-4,0	-2,4	+ 9,60	.	16,00	5,76	-0,45	-0,20
1891	44,3	436,7	+7,3	+10,4	+75,92	.	53,29	108,16	+0,81	+0,88
1892	41,8	380,7	-2,6	-6,3	+16,38	.	6,76	39,69	-0,29	-0,53
1893	40,5	430,5	-6,9	+9,0	.	-62,10	47,61	81,00	-0,77	+0,76
1894	49,5	422,1	+12,2	+12,5	+152,50	.	148,84	156,25	+1,36	+1,06
1895	41,5	304,6	-5,7	-17,7	+100,89	.	32,49	313,29	-0,64	-1,50
1896	47,2	337,9	+5,6	-6,7	.	-37,52	31,36	44,89	+0,62	-0,57
1897	41,5	355,8	-7,4	+3,3	.	-24,42	54,76	10,89	-0,82	+0,28
1898	43,8	389,9	-1,8	+19,3	.	-34,74	3,24	372,49	-0,20	+1,63
1899	50,1	333,3	+14,1	+4,9	+69,09	.	198,81	24,01	+1,57	+0,41
1900	40,4	216,9	-13,5	-28,5	+384,75	.	182,25	812,25	-1,50	-2,41
1901	43,8	292,4	-10,5	+4,8	.	-50,40	110,25	23,04	-1,17	+0,41
1902	55,5	283,3	+12,8	+3,1	+39,68	.	163,84	9,61	+1,43	+0,26
1903	54,8	269,9	+6,6	-9,7	.	-64,02	43,56	94,09	+0,74	-0,82
1904	51,8	311,4	-7,5	-4,2	+31,50	.	56,25	17,64	-0,84	-0,36
1905	51,2	337,5	-13,1	-4,7	+61,57	.	171,61	22,09	-1,46	-0,40
1906	66,9	422,1	+10,9	+10,2	+111,18	.	118,81	104,04	+1,22	+0,86
1907	70,3	430,5	+12,4	+9,2	+114,08	.	153,76	84,64	+1,38	+0,78
1908	61,5	414,0	-5,7	+1,2	.	-6,84	32,49	1,44	-0,64	+0,10
1909	63,2	368,3	-5,3	-16,0	+84,80	.	28,09	256,00	-0,59	-1,35
1910	64,3	409,7	-1,3	-14,6	+18,98	.	1,69	213,16	-0,14	-1,23
1911	74,5	569,7	+12,3	+10,6	+130,38	.	151,29	112,36	+1,37	+0,90
1912	62,4	637,9	-6,5	+17,6	.	-114,40	42,25	309,76	-0,72	+1,49
1913	67,4	590,0	.	.	.	.	.	.	.	.
1914	65,3	504,2	.	.	.	.	.	.	.	.
	1411,5	**10490,4**			**+1401,30**	**-394,44**	**1849,30**	**3216,55**		

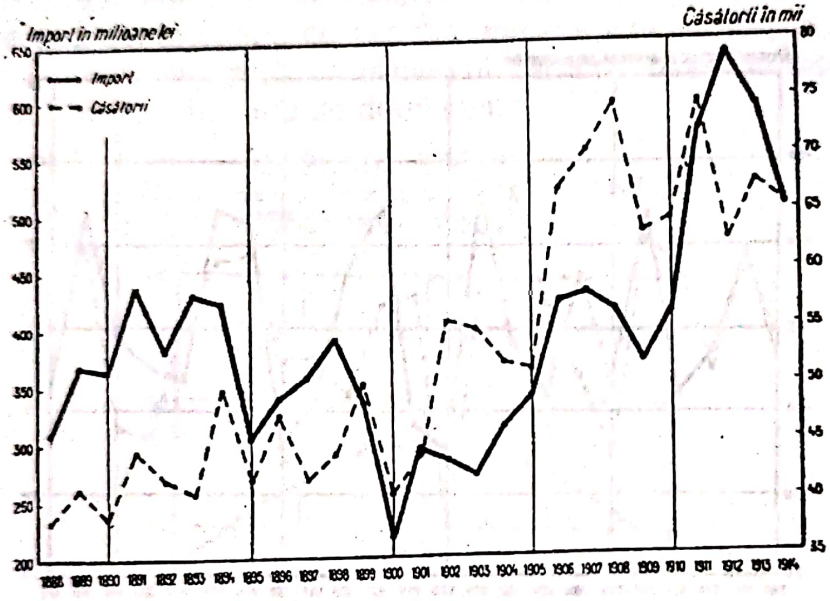
*Extras din **Anuarele statistice ale României** pe anii 1904, 1915/1916.*

Calculul coeficientului de covariație diferențială se face pornind de la aceste valori, fără să întâmpinăm vreo dificultate specială: $k = 0,46$.

Scanned with OKEN Scanner

În ceea ce privește calculul celuilalt coeficient de covariație, trebuie să observăm că nu putem admite ca valoarea teoretică a mediei aritmetice a componentelor reziduale să aibă o valoare diferită de zero. Căci dacă acest din urmă caz s-ar întâmpla, ar însemna că am comis o eroare *sistematică* în calculul tendinței de lungă durată și că aceasta din urmă ar trebui să fie sau coborâtă sau urcată, astfel ca media componentelor reziduale să fie nulă. Evident, *valoarea statistică* a mediei componentelor reziduale poate avea totuși o valoare oarecare, considerațiile precedente privind numai media teoretică. Din această cauză coeficientul de covariație tangențială al unor serii formate din componente reziduale se calculează direct prin formula:

$$c = \frac{\sum r_x r_y}{\sqrt{\sum r_x^2 \sum r_y^2}}. \qquad \text{(155 bis)}$$



Tabelul 153

Graficul 70. Comparația între mișcările în timp ale importului în valoare și ale numărului de căsătorii celebrate (România, 1888 - 1914).

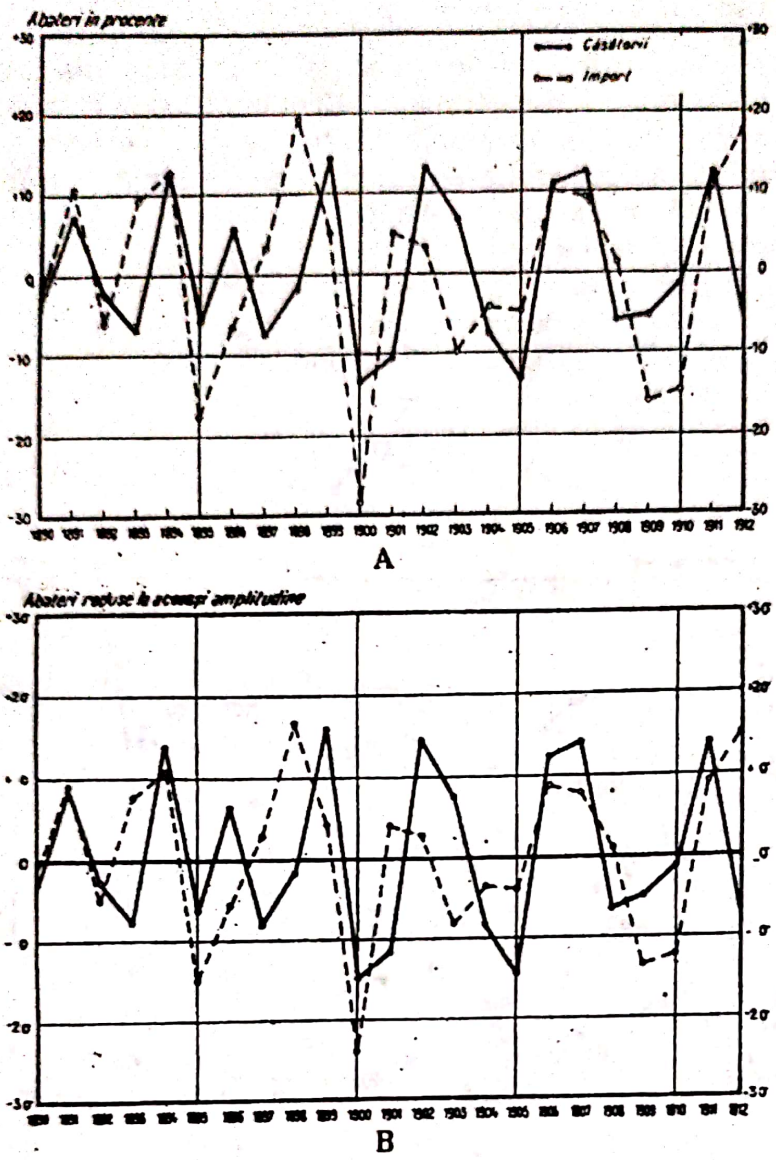
Aplicând această formulă, obținem, în cazul de față:

$$c = \frac{1401{,}30 - 394{,}44}{\sqrt{1849{,}30 \times 3216{,}55}} = \frac{1006{,}86}{2438{,}93} = 0{,}41\ .$$

Aceste valori ne indică o covariație certă între nupțialitate și importul în valoare[1] (graficul 71).

[1] *R. H. Hooker în* ***On the Correlation of the Mariage-rate with Trade*** *(Journal of the Royal Statistical Society, 1901, p. 485) stabilește pentru Anglia o covariație mult mai puternică între nupțialitate și comerțul exterior (c = 0,86). Diferența poate să provină fie dintr-o structură socială diferită, fie din cauză că metoda lui Hooker este mai rafinată decât cea pe care noi o întrebuințăm doar cu titlu de exemplificare a unui procedeu de investigație.*

Scanned with OKEN Scanner



Tabelul 153

Graficul 71. Covariaţia mişcărilor ondulatorii ale importului în valoare şi numărului de căsătorii celebrate.

193. *Covariaţie cu întârziere*. Unul dintre cele mai utile cazuri sub care se poate prezenta în practică covariaţia - util pentru că alcătuieşte pivotul mai tuturor procedeelor de previziune economică - este cel al covariaţiei cu întârziere.

Seriile în timp a două fenomene economice pot să nu conducă la nici o covariaţie apreciabilă dacă fiecare termen este comparat cu cel corespunzător aceleiaşi epoci; însă aceleaşi serii pot să apară într-o strânsă covariaţie dacă recurgem la o comparaţie dintre termenii separaţi printr-un oarecare interval de timp. În acest din urmă caz, înseamnă că o variaţie a primului fenomen nu va influenţa imediat mersul celui de-al doilea fenomen, ci variaţia corespunzătoare acestuia din urmă va avea loc ceva mai târziu.

Scanned with OKEN Scanner

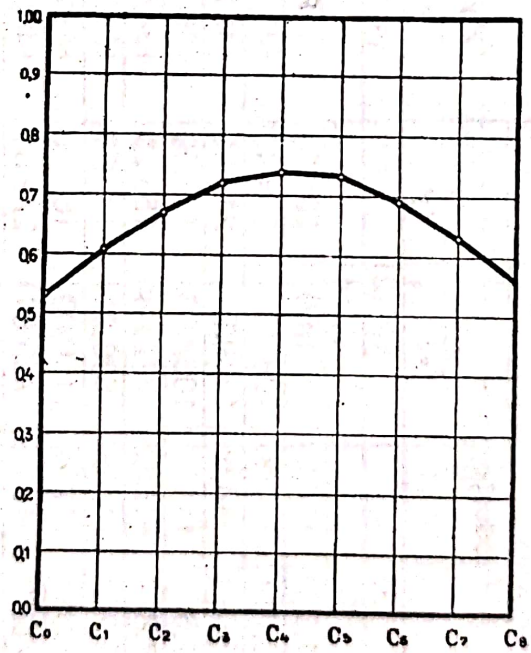
Covariaţia cu întârziere poate să fie recunoscută şi prin intermediul reprezentării grafice: una din serii fiind reprezentată pe hârtie de calc, o suprapunem peste cealaltă şi, printr-o mişcare de translaţie de-a lungul axei timpului, vedem cum covariaţia celor două curbe - slabă sau inexistentă la început - se accentuează, atinge o valoare maximă, pentru a deveni în urmă din ce în ce mai ştearsă.

Cu toate acestea, instrumentul numeric este, şi în acest caz, de preferat. Să ne referim la seriile din tabelul 154. Coeficientul de covariaţie tangenţială obişnuit va fi:

$$c_0 = \frac{10947,39}{\sqrt{73105,25 \times 5872,53}} = 0,53.$$

Deplasând acum întreaga serie a producţiei astfel ca fiecare termen al acesteia să fie aşezat cu două luni mai în urmă, adică să punem în legătură valoarea coeficientului de plasare din ianuarie cu aceea a indicelui de producţie din martie, pe cea din februarie cu cea din aprilie ş.a.m.d., vom obţine, pornind de la această corespondenţă, coeficientul de covariaţie tangenţială dintre plasare şi producţie cu întârziere de două luni:

$$c_2 = \frac{13598,15 - 746,26}{\sqrt{66569,36 \times 5466,40}} = 0,67\ ^{1}.$$



Graficul 72. Coeficienţii de covariaţie cu întârziere.

[1] *La numitor nu vom mai avea aceleaşi valori ca în cazul lui c_0 din cauză că c_2 este obţinut din aceleaşi serii din care lipsesc însă primii doi termeni r_y şi ultimii doi r_x:*

73105,25 - 3422,25 - 3113,64 = 66569,36
5872,53 - 201,64 - 204,49 = 5466,40

Scanned with OKEN Scanner

TABELUL 154

Covariația cu întârziere dintre coeficientul de plasare al mâinii de lucru și indicele de producție

(Franța 1924 - 1927)[1]

LUNA	Componente reziduale		Pătrate		Produse	Produse $r_x r_y$ cu întârziere de			
	Plasare r_x	Producție r_y	r_x^2	r_y^2	$r_x r_y$	2 luni	4 luni	6 luni	8 luni
(1)	(2)		(3)		(4)	(5)			
1924									
Ianuarie	-1,3	+14,2	1,69	201,64	-18,46	-18,85	-16,25	-16,51	-21,97
Februarie	+33,0	+14,3	1089,00	204,49	+471,90	+372,90	+382,80	+491,70	+597,30
Martie	+33,9	+14,5	1149,21	210,25	+491,55	+423,75	+430,53	+572,91	+444,09
Aprilie	+32,6	+11,3	1062,76	127,69	+368,38	+378,16	+485,78	+590,06	+430,32
Mai	+27,8	+12,5	772,84	156,25	+347,50	+353,06	+469,82	+364,18	+286,34
Iunie	+23,2	+11,6	538,24	134,56	+269,12	+345,68	+419,92	+306,24	+171,68
Iulie	+34,0	+12,7	1156,00	161,29	+431,80	+574,60	+445,40	+350,20	+190,40
August	+27,2	+14,9	739,84	222,01	+405,28	+492,32	+359,04	+201,28	+106,08
Septembrie	+14,0	+16,9	196,00	285,61	+236,60	+183,40	+144,20	+78,40	+30,80
Octombrie	+11,8	+18,1	139,24	327,61	+213,58	+155,76	+87,32	+46,02	+4,72
Noiembrie	+9,7	+13,1	94,09	171,61	+127,07	+99,91	+54,32	+21,34	-2,91
Decembrie	+10,8	+13,2	116,64	174,24	+142,56	+79,92	+42,12	+4,32	-11,88

[1] *Pentru a nu complica tabelul, publicăm numai componentele reziduale ale fenomenelor considerate. Tendințele de lungă durată au fost calculate pe baza mediilor anuale (1920 - 1929), prin interpolare liniară, și sunt date de ecuațiile: $Y = 95,3 + 0,91t$ pentru plasament și $Y = 100,2 + 0,97t$ pentru producție.*

Scanned with OKEN Scanner

(continuare tab. 154)

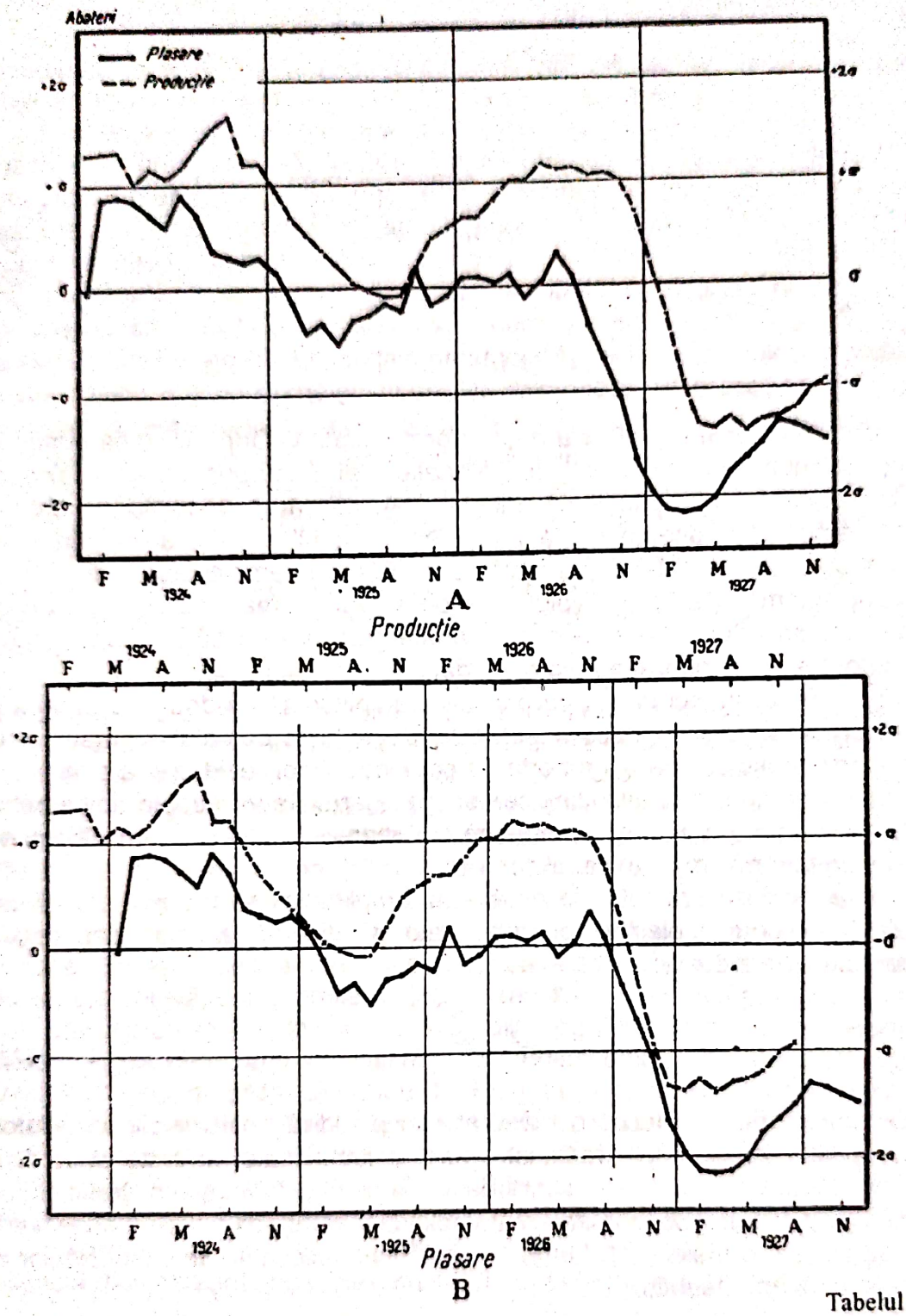
(1)	(2)		(3)		(4)	(5)			
1925									
Ianuarie	+5,5	+10,3	30,25	106,09	+56,65	+30,80	+12,10	-1,65	-4,95
Februarie	-4,8	+7,4	23,04	54,76	-35,52	-18,72	-1,92	+5,28	-10,08
Martie	-16,9	+5,6	285,61	31,36	-94,64	-37,18	+5,07	+15,21	-86,19
Aprilie	-13,6	+3,9	184,96	15,21	-53,04	-5,44	+14,96	-28,56	-84,32
Mai	-20,6	+2,2	424,36	4,84	-45,32	+6,18	+18,54	-105,06	-148,32
Iunie	-12,4	+0,4	153,76	0,16	-4,96	+13,64	-26,04	-76,88	-91,76
Iulie	-10,1	-0,3	102,01	0,09	+3,03	+9,09	-51,51	-72,72	-93,93
August	-6,0	-1,1	36,00	1,11	+6,60	-12,60	-37,20	-44,40	-67,20
Septembrie	-8,8	-0,9	77,44	0,81	+7,92	-44,88	-63,36	-81,84	-99,44
Octombrie	+6,8	+2,1	46,24	4,41	+14,28	+42,16	+50,32	+76,16	+89,08
Noiembrie	-6,5	+5,1	42,25	26,01	-33,15	-46,80	-60,45	-73,45	-79,95
Decembrie	-3,5	+6,2	12,25	38,44	-21,70	-25,90	-39,20	-45,85	-43,40
1926									
Ianuarie	+2,6	+7,2	6,76	51,84	+18,72	+24,18	+29,38	+31,98	+30,42
Februarie	+3,2	+7,4	10,24	54,76	+23,68	+35,84	+41,92	+39,68	+33,08
Martie	+1,5	+9,3	2,25	86,49	+13,95	+16,95	+18,45	+17,55	+16,65
Aprilie	+4,2	+11,2	17,64	125,44	+47,04	+55,02	+52,08	+49,98	+32,76
Mai	-4,8	+11,3	23,04	127,69	+54,24	-59,04	-56,16	-53,28	-10,08
Iunie	+0,7	+13,1	0,49	171,61	+9,17	+8,62	+8,33	+5,46	-2,52
Iulie	+11,5	+12,3	132,25	151,29	+141,45	+134,55	+127,65	+24,15	-115,00
August	+1,8	+12,4	3,24	153,76	+22,32	+21,42	+14,04	-6,48	-26,46

Scanned with OKEN Scanner

(continuare tab. 154)

(1)	(2)		(3)		(4)	(5)			
Septembrie	-15,7	+11,7	246,49	136,89	-183,69	-174,27	-32,97	+157,00	+240,21
Octombrie	-27,7	+11,9	767,29	141,61	-329,63	-216,06	+99,72	+407,19	+393,34
Noiembrie	-41,2	+11,1	1 697,44	123,21	-457,32	-86,52	+412,00	+630,36	+638,60
Decembrie	-64,9	+7,8	4 212,01	60,84	-506,22	+233,64	+954,03	+921,58	+934,56
1927									
Ianuarie	-76,2	+2,1	5 806,44	4,41	-160,02	+762,00	+1 165,86	+1 181,10	+1 082,04
Februarie	-83,1	-3,6	6 905,61	12,96	+299,16	+1 221,57	+1 180,02	+1 196,64	+1 096,92
Martie	-84,9	-10,0	7 208,01	100,00	+849,00	+1 298,97	+1 315,95	+1 205,58	+959,37
Aprilie	-83,4	-14,7	6 955,56	216,09	+1 225,98	+1 184,28	+1 200,96	+1 100,88	+859,02
Mai	-79,4	-15,3	6 304,36	234,09	+1 214,82	+1 230,70	+1 127,48	+897,22	-
Iunie	-69,7	-14,2	4 858,09	201,64	+989,74	+1 003,68	+920,04	+717,91	-
Iulie	-64,3	-15,5	4 134,49	240,25	+996,65	+913,06	+726,59	-	-
August	-58,8	-14,4	3 457,44	207,36	+846,72	+776,16	+605,66	-	-
Septembrie	-51,1	-14,2	2 611,21	201,64	+725,62	+577,43	-	-	-
Octombrie	-52,3	-13,2	2 735,29	174,24	+690,36	+538,69	-	-	-
Noiembrie	-55,8	-11,3	3 113,64	127,69	+630,54	-	-	-	-
Decembrie	-58,5	-10,3	3 422,25	106,09	+602,55	-	-	-	-
			73 105,25	**5 872,53**	**-1 997,91**	**-746,26**	**-385,06**	**-606,68**	**-1 000,36**
					+12 945,30	**+13 598,15**	**+13 422,38**	**+11 707,56**	**+8 667,98**

Scanned with OKEN Scanner



Tabelul 154

Graficul 73. Covariația cu întârziere dintre plasare și producție (Franța, 1924 - 1927).

Scanned with OKEN Scanner

La fel, obținem:

$$c_1 = 0{,}61 \quad c_3 = 0{,}72 \quad c_4 = 0{,}74 \quad c_5 = 0{,}73$$
$$c_6 = 0{,}69 \quad c_7 = 0{,}63 \quad c_8 = 0{,}56.$$

Variația acestui șir de coeficienți (graficul 72) ne pune în evidență existența unei covariații cu întârziere între fenomenele considerate. Intervalul optim, indicat de coeficientul de covariație care are valoarea cea mai mare, este în cazul de față egal cu 4 luni, ceea ce înseamnă că variațiile coeficientului de plasare preced pe cele ale indicelui de producție cu acest interval de timp (graficul 73 A și B). Primul fenomen poate fi considerat deci ca un element anunțător al celui de-al doilea. Majoritatea sistemelor de previziune cunoscute sub numele de barometre economice utilizează covariația cu întârziere.

194. *Comparația grafică a mișcărilor ondulatorii.* După cum ne putem da seama comparând ordinul de variație al componentelor reziduale corespunzătoare fenomenelor din tabelul 154, mișcarea ondulatorie poate să prezinte amplitudini diferite de la un fenomen la altul. Sunt adică fenomene oarecum stabile, prezentând, prin urmare, ondulații de amplitudine redusă, după cum sunt fenomene a căror evoluție cunoaște variații mai pronunțate. Această diferență de amplitudini face ca o comparație grafică a mișcărilor ondulatorii să nu fie întotdeauna o operație prea comodă.

S-a căutat atunci un mijloc prin care amplitudinile a două mișcări să fie făcute comparabile chiar dacă acestea se deosebesc complet ca mărime. Un mijloc foarte firesc constă din raportarea componentelor reziduale la o măsură a ordinului de mărime a amplitudinii corespunzătoare. Se va adopta adică pentru aceste componente o unitate de măsură mai mare în cazul unei amplitudini mai mari și o unitate mai mică în cazul contrar.

Determinarea ordinului de mărime al amplitudinii se reduce la problema reprezentării unui colectiv printr-o valoare tipică. Media pătratică a componentelor reziduale:

$$\sigma^2 = \frac{\sum r^2}{n} \qquad (156)$$

s-a consacrat prin tradiție ca o măsură a amplitudinii unei mișcări ondulatorii. Multă vreme componentele reziduale au fost identificate cu niște *abateri*, iar (156) cu o *abatere-tip*. Acest punct de vedere este bineînțeles greșit; totuși, utilitatea mediei pătratice ca măsură a amplitudinii rămâne un procedeu legitim.

Aplicarea formulei (156) nu întâmpină nici o dificultate. În cazul seriilor din tabelul 153, avem respectiv:

$$\sigma_1 = \sqrt{\frac{1849{,}30}{23}} = 8{,}97 \qquad \sigma_2 = \sqrt{\frac{3216{,}55}{23}} = 11{,}83.$$

Scanned with OKEN Scanner

O dată σ obținut, se trece la calculul componentelor reduse, $\frac{r}{\sigma}$, care au avantajul de a prezenta aceeași amplitudine, oricare ar fi seria considerată (tabelul 153, coloana 7).

Graficele 71 B și 73 sunt executate pe baza acestor principii.

195. *Regresia în timp*. Problema exprimării numerice a unei covariații se pune în mod firesc spiritului nostru, imediat ce existența acestei covariații a fost concepută, explicată și verificată. Chestiunea prezintă o analogie formală cu regresia dintre atribute.

Nu vom mai insista asupra punctelor ce deosebesc fundamental pe aceasta din urmă de problema regresiei în timp, care are ca scop tot determinarea relației funcționale ce leagă două fenomene, dedusă însă prin intermediul comparației evoluțiilor lor. Vom trece la examinarea unui caz concret, ocupându-ne și de data aceasta numai de cazul regresiei liniare. Formulele sunt identice cu cele introduse în cap. XI și XII.

Problema de care vrem să ne ocupăm pentru a ilustra procedeele de calcul ale regresiei în timp este aceea a influenței elementelor meteorologice asupra producției agricole medii. Vom analiza această problemă în modul cel mai amănunțit pe care ni-l permite cadrul acestui curs; tocmai pentru ca să oferim un exemplu cât mai viu al uneia dintre cele mai frecvente aplicații ale statisticii matematice.

Existența unei influențe a elementelor meteorologice asupra producției agricole este o chestiune pe care ar fi absurd să o mai discutăm. Problema vrea însă să determinăm care anume din fenomenele meteorologice constituie factori esențiali de înrâurire asupra recoltelor. Evident că dintr-un început trebuie să ne referim la o regiune *omogenă* din punct de vedere *climatic*. Nu am ajunge la nici un rezultat înglobând în cercetările noastre dintr-o dată două sau mai multe regiuni diferite din punct de vedere al normalelor meteorologice, căci în timp ce pentru Bărăgan - regiune întrucâtva secetoasă - orice an cu o cantitate de ploaie superioară normalei conduce în general la o recoltă deasupra mediei, pentru o regiune ploioasă - cum ar fi regiunea deluroasă a Moldovei - excedentul de ploaie poate, dimpotrivă, compromite definitiv o recoltă. Vedem acum mai clar cum trebuie privită problema: este vorba de a stabili o legătură dintre *abaterile elementelor meteorologice de la normala unei astfel de regiuni omogene și producția agricolă*[1]. Dacă vom găsi, de pildă, că pentru Bărăgan nu există covariație netă între recolta de porumb și cantitatea de ploaie căzută în iulie, aceasta nu va însemna că, plouă sau nu în această lună, recolta nu se va resimți. O astfel de constatare va trebui privită în sensul că *frecvența ploilor în Bărăgan și în cursul lunii iulie prezintă o stabilitate suficientă și necesară culturii porumbului în acea regiune*.

Evident că printre elementele meteorologice care ne interesează în această direcție sunt unele care nu pot fi decât foarte greu introduse în calcul:

---

[1] *Omogenitatea unei regiuni din punct de vedere climatic este o altă problemă destul de complicată. Nu ne oprim asupra ei. Menționăm doar că procedeele descrise în cap. IX sunt primele la care trebuie să facem apel pentru tratarea acestei probleme.*

Scanned with OKEN Scanner

Primul îngheţ general s-a produs înaintea primei zăpezi sau nu? După dezgheţul din primăvară a mai intervenit sau nu un îngheţ trecător? O cantitate de ploaie căzută în cursul unei luni a provenit din ploi frecvente şi de scurtă durată sau dintr-un număr mai mic de ploi mai intense?

Iată atâtea chestiuni pe care forţat nu le putem considera numai cu posibilităţile pe care ni le pun la dispoziţie statisticile noastre de azi. Acest lucru constituie o piedică serioasă în calea oricărei încercări de a stabili o covariaţie între fenomenele meteorologice şi producţia grâului din România, ştiut fiind că la noi se cultivă de preferinţă grâul de toamnă. Din această cauză, ne vom referi aici la cazul porumbului care are o vegetaţie mai scurtă şi care - lucru esenţial pentru metodă - nu intră în nici una din lunile de îngheţ.

Limitându-ne la judeţele Brăila şi Ialomiţa, a căror suprafaţă constituie aproape în întregime o regiune omogenă - Bărăganul -, ne vom pune problema determinării lunilor caracteristice din punctul de vedere al ploii pentru cultura porumbului. Pe baza unor calcule pe care nu le reproducem pentru că nu prezintă nici un interes scopului nostru din momentul de faţă, am stabilit că pentru perioada 1895 - 1915 cele mai pronunţate covariaţii dintre cantitatea de ploaie căzută şi producţia medie la hectar sunt din lunile iunie şi septembrie:

$$c_{12} = +0{,}601 \qquad c_{13} = -0{,}432$$

unde indicele 1 corespunde producţiei medii, iar 2 şi 3, respectiv cantităţilor de ploaie căzute în cele două luni menţionate. Rămânând numai la aceste două caracteristici meteorologice, va trebui să calculăm în prealabil coeficienţii ecuaţiei de regresie (128):

$$x_{1\cdot 23} = b_{12\cdot 3}x_2 + b_{13\cdot 2}x_3,$$

unde x înseamnă în orice caz o abatere de la medie sau de la normală. Pentru producţie vom adopta ca normală media aritmetică mobilă pe cinci ani[1], iar pentru cantitatea de ploaie, media aritmetică a şirului[2]. Amănuntele de calcul sunt expuse în tabelul 155.

Coeficienţii de covariaţie totală se stabilesc prin formula (155):

---

[1] *Media producţiei pe cinci ani se impune în mod natural ca valoare normală. Nu poate fi vorba de a căuta o altă normală printr-o interpolare liniară, deoarece însăşi evoluţia producţiei la hectar nu pare să se poată încadra într-o formă matematică, iar pe deasupra ne găsim în faţa unui fenomen cu ereditate: producţia medie dintr-un an este în bună parte legată de cea a anului precedent şi influenţează la rându-i pe cea a anului următor. Fenomenul este destul de complex şi, după câte ştiu, nu s-a făcut nici un studiu complet în această direcţie.*

[2] *Cantitatea de ploaie este de fapt o medie aritmetică calculată pe baza observaţiilor lunare din centrele de observaţie mai importante din Bărăgan. Datele, atât pentru producţie, cât şi pentru ploi, au fost culese din diferite publicaţii (unele din* ***Monitorul Oficial*** *chiar), aşa că nu vom da o listă exactă a lor.*

Scanned with OKEN Scanner

$$c_{12} = \frac{31848{,}35 - 2156{,}55}{\sqrt{54119{,}97 \times 45095{,}57}} = 0{,}601,$$

$$c_{13} = \frac{7144{,}08 - 40412{,}54}{\sqrt{54119{,}97 \times 109666{,}38}} = -0{,}432\ ,$$

$$c_{23} = \frac{13126{,}06 - 36717{,}68}{\sqrt{109666{,}38 \times 45095{,}57}} = -0{,}335\ ,$$

iar cei de covariație parțială și multiplă, prin formule identice cu cele întrebuințate în cazul corelației (125):

$$c_{12\cdot3} = \frac{c_{12} - c_{13}c_{23}}{\sqrt{1-c_{13}^2}\sqrt{1-c_{23}^2}} = 0{,}534 \quad c_{13\cdot2} = \frac{c_{13} - c_{12}c_{23}}{\sqrt{1-c_{12}^2}\sqrt{1-c_{23}^2}} = -0{,}306$$

$$c_{1\cdot23}^2 = \frac{c_{12}^2 + c_{13}^2 - 2c_{12}c_{23}c_{31}}{1-c_{23}^2} = 0{,}421134 \quad C_{1\cdot23} = 0{,}649.$$

La fel:

$$b_{12\cdot3} = 0{,}563 \qquad b_{13\cdot2} = -0{,}183$$

și deci ecuația căutată va fi:

$$x_{1\cdot23} = 0{,}563x_2 - 0{,}183x_3.$$

Precizia determinării recoltei prin această ecuație este dată de formula cunoscută:

$$\sigma_{1\cdot23} = \sigma_1\sqrt{1-c_{12}^2}\sqrt{1-c_{13\cdot2}^2} = 42{,}92.$$

Graficul 74 ne permite o comparație între valorile statistice ale recoltei de porumb și cele deduse din cunoașterea cantităților de ploaie căzută în lunile iunie și septembrie (ultima coloană a tabelului 155). Pe fiecare ordonată am purtat de o parte și de alta a valorii teoretice segmente egale cu $2\sigma_{1\cdot23}$ pentru a forma regiunea de variație statistică a fenomenului (§ 94).

Ținând seama de imprecizia datelor relative la producția medie, destul de importantă în cazul unui singur județ, de faptul că pentru ploaie nu am utilizat decât observațiile din trei stațiuni (exactitatea acestora fiind destul de dubioasă și ea), precum și de acela că nu am făcut să intervină în calculul nostru elemente mai indicate decât cantitatea de ploaie căzută în cursul unei luni, concordanța între valorile statistice și cele calculate este extrem de satisfăcătoare. Printr-un procedeu asemănător s-a ajuns în Statele Unite să se prevadă recolta medie de porumb cu câteva luni înainte de epoca culesului, datele astfel evaluate întrecând cu mult în precizie chiar pe cele stabilite pe baza rapoartelor experților agronomi din diferite regiuni.

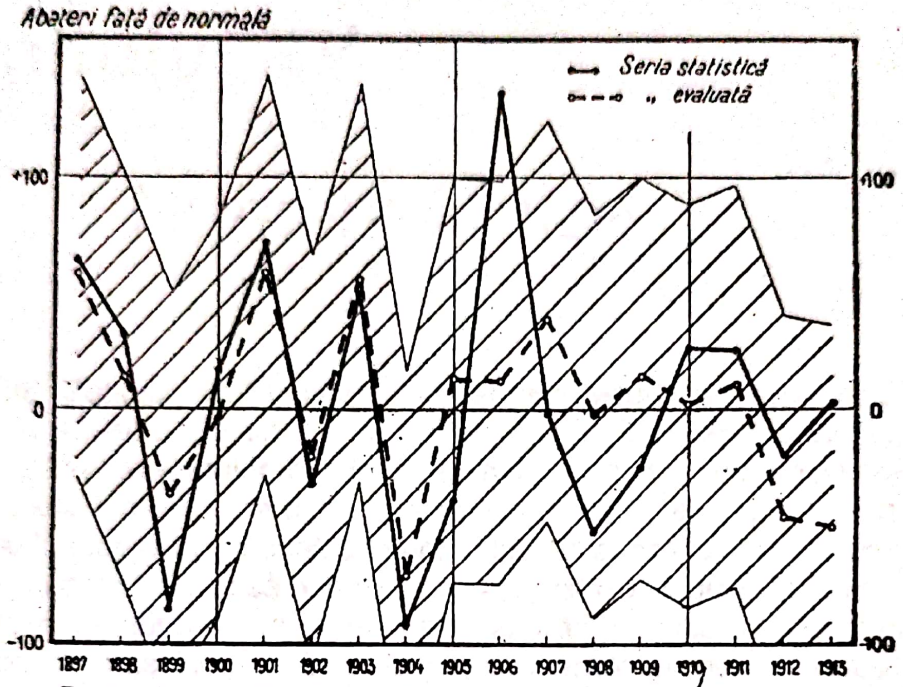
Scanned with OKEN Scanner

TABELUL 155

Calculul regresiei în timp dintre producția medie de porumb și cantitățile de ploaie căzute în iunie și septembrie

Anii	Producția medie la ha hectolitri)	Cantitatea de ploaie		Abateri față de normală			Pătrate			Produse			Evaluări prin regresie	
		Iunie	Sept	x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3	$x_{1\cdot 2}$	$x_{1\cdot 23}$
1895	14,9	256	160											
1896	8,4	574	214											
1897	20,7	1295	254	+63,2	+88,8	-43,5	3 994,24	7 885,44	1 892,25	+5 612,16	-2 749,20	-3 862,80	+58,4	+58,0
1898	16,9	692	141	+30,8	+0,9	-68,6	948,64	0,81	4 705,96	+27,72	-2 112,88	-61,74	+0,6	+13,1
1899	2,5	297	564	-84,4	-56,7	+25,5	7 123,36	3 214,89	650,25	+4 785,48	-2 152,20	-1 445,85	-37,3	-36,6
1900	16,1	537	255	+16,8	-21,7	-43,3	282,24	470,89	1 874,89	-364,56	-727,44	+939,61	-14,3	-4,3
1901	24,0	1 338	347	+70,9	+95,1	-22,8	5 026,81	9 044,01	519,84	+6 742,59	-1 616,52	-2 168,28	+62,6	+57,7
1902	9,4	343	259	-31,7	-50,0	-42,4	1 004,89	2 500,00	1 797,76	+1 585,00	+1 344,08	+2 120,00	-32,9	-20,4
1903	18,2	1 120	0	+48,9	+64,6	-100,0	2 391,21	4 173,16	10 000,00	+3 158,94	-4 890,00	-6 460,00	+42,5	+54,7
1904	1,1	105	1 009	-91,6	-84,7	+124,5	8 390,56	7 174,09	15 500,25	+7 758,52	-11 404,20	-10 545,15	-55,7	-70,5
1905	8,4	696	160	-39,7	+1,5	-64,4	1 576,09	2,25	4 147,36	-59,55	+2 556,68	-96,60	+1,0	+12,6
1906	28,1	683	160	+135,7	-0,4	-64,4	18 414,49	0,16	4 147,36	-54,28	-8 739,08	+25,76	-0,3	+11,6
1907	13,9	1 008	166	-1,8	+47,0	-63,1	3,24	2 209,00	3 981,61	-84,60	+113,58	-2 965,70	+30,9	+38,0
1908	8,1	682	507	-51,9	-0,6	+12,8	2 693,61	0,36	163,84	+31,14	-664,32	-7,68	-0,4	-2,6
1909	12,3	1 106	957	-24,6	+61,3	+112,9	605,16	3 757,69	12 746,41	-1 507,98	-2 777,34	+6 920,77	+40,3	+13,8
1910	21,8	823	685	+26,0	+20,0	+52,4	676,00	400,00	2 745,76	+520,00	+1 362,40	+1 048,00	+13,2	+1,7
1911	25,5	930	711	+24,5	+35,6	+58,2	600,25	1 267,36	3 387,24	+872,20	+1 425,90	+2 071,92	+23,4	+9,3
1912	18,8	422	1 041	-19,6	-38,5	+131,6	384,16	1 482,25	17 318,56	+754,60	-2 579,36	-3 066,60	-25,3	-45,8
1913	24,0	419	1 147	+2,2	-38,9	+155,2	4,84	1 513,21	24 087,04	-85,58	+341,44	-6 037,28	-25,6	-50,3
1914	26,8	853	509											
1915	22,3	215	192											
		14 403	**9 438**				**54 119,79**	**45 095,57**	**109 666,38**	**+31 848,35**	**+7 144,08**	**+13 126,06**		
										-2 156,55	**-40 412,54**	**-36 717,68**		

Scanned with OKEN Scanner



Partea haşurată reprezintă regiunea de variație statistică.

Tabelul 155

Graficul 74. Comparație între producția medie de porumb și valorile prevăzute prin regresia în timp.

Scanned with OKEN Scanner

Capitolul XV

Numere indice

196. *Generalităţi.* Sunt cazuri în care interesul nostru poate să fie atras de variaţia în ansamblu a unui grup de fenomene, cum ar fi preţurile, producţia, salariile, cursurile valorilor mobiliare etc. Ne dăm seama că, deşi fiecare element dintr-un astfel de grup are o mişcare în timp proprie, există totuşi un factor care acţionează asupra lor în acelaşi timp; influenţa acestuia vrem să o desprindem din întregul ansamblu şi să o individualizăm printr-o singură valoare.

Instrumentul care ne permite să măsurăm variaţia relativă a unui grup de fenomene considerate în ansamblu poartă numele de *număr indice*[1]. Numerele indice pot să fie construite pentru scopuri cât mai diferite şi pot oglindi, după caz, fie o variaţie în timp, fie una în spaţiu.

Determinarea formulei pe baza căreia să se calculeze un număr indice pare, la prima vedere, să se confunde cu problema reprezentării unui colectiv printr-o valoare tipică. Faptul că această valoare trebuie să servească de bază comparaţiei relative în timp şi spaţiu şi că sensul comparaţiei este un element perfect reversibil vine să complice problema într-un astfel de mod încât, cu toate studiile întreprinse în această direcţie de cei mai de seamă statisticieni, soluţia ei, cel puţin în mod teoretic, nu este încă pe deplin stabilită. În practică însă, rezultatele sunt prea puţin influenţate de structura matematică a numărului indice, atâta vreme cât aceasta este aleasă dintre acelea asupra cărora statisticienii sunt mai mult sau mai puţin de acord.

Problema se izbeşte însă în aplicaţiile concrete de alte dificultăţi, mai puţin academice, însă mult mai complicate şi care pot influenţa în mod considerabil utilitatea indicelui calculat. Dintre acestea, compoziţia indicelui este una din cele mai importante. Să ne referim la cazul unui indice al variaţiei preţurilor. Lăsând la o parte faptul că un astfel de indice poate să fie calculat pentru scopuri diferite, cum ar fi, de pildă, indicele preţurilor de detaliu, de gros, al preţurilor produselor indigene, importate, agricole etc., prima chestiune care se ridică este aceea a numărului de articole care vor intra în compoziţia lor, a naturii lor, precum şi a calităţii fiecărui articol în parte.

Statistica Generală a Franţei calculează un indice al preţurilor de gros pe baza a 45 de articole, în timp ce în Statele Unite, *Bureau of Labor Statistics*

---

[1] Index numbers, *în engleză;* nombres indices, *în franceză.*

Scanned with OKEN Scanner

utilizează 550, iar *War Industries Board* a construit un indice conținând 1366 articole.

Ar fi greu de fixat chiar numai niște limite inferioare și superioare pentru numărul articolelor care să intre în compoziția unui indice de prețuri și care să fie valabile pentru orice scop și orice țară.

Alegerea articolelor care să intre în compoziția indicelui va conduce, dacă este făcută după anumite principii, în mod implicit, la o determinare a numărului lor. În general, mai multe articole vor conduce la un indice mai reprezentativ. Nu trebuie însă uitat că înseși articolele trebuie să fie și ele reprezentative la rândul lor, adică nu vom alege decât articole ale căror prețuri nu sunt prea legate între ele. O independență cât mai mare între prețurile articolelor utilizate va asigura o structură echilibrată a indicelui și ne va feri de o ponderare ascunsă, dând unui grup de articole o importanță mai mare decât este natural să-i fie acordată. Prin aplicarea principiilor de mai sus, limităm și superior, și inferior numărul articolelor din compoziția indicelui.

Calitatea fiecărui articol în parte, precum și condițiile de livrare în cazul prețurilor de gros trebuie să fie precizate cu cea mai mare grijă pentru ca astfel indicele să nu fie influențat de alte variații decât acelea pe care le urmărim.

Considerații similare se pot face și în legătură cu numerele indice pentru producție, salarii etc. Dacă ne vom ocupa mai ales de indicii variației prețurilor este din cauză că aceștia au atras de la început atenția statisticienilor și economiștilor, precum și pentru faptul că dau naștere la cele mai importante aplicații.

197. *Indici bugetari.* De la început să spunem că indicii de preț sunt de două feluri:

a) *indici bugetari*;

b) *indici monetari.*

Indicii bugetari se construiesc cu scopul de a ne pune în evidență variațiile unui buget oarecare de schimb de mărfuri datorate mișcării prețurilor. Printre indicii bugetari putem cita, în primul rând, ca cel mai cunoscut, indicele care reflectă variațiile costului vieții. Acest indice pune în evidență variațiile sumei de bani pe care un individ mediu dintr-o anumită localitate și anumită clasă socială trebuie să o cheltuiască pentru ca să-și mențină existența.

Pentru această categorie de indici problema alegerii unei formule de calcul nici nu se pune, aceasta din urmă fiind indicată de însăși definiția indicelui bugetar. Astfel dacă $q_1, q_2, \ldots, q_n$ reprezintă cantitățile fixate în prealabil prin bugetul a cărui variație o urmărim și $p_1^0, p_2^0, \ldots, p_n^0$, prețurile articolelor ce alcătuiesc acest buget, suma necesară cumpărăturilor va fi:

$$S_0 = q_1 p_1^0 + q_2 p_2^0 + \ldots + q_n p_n^0 = \sum q p^0 .$$

Aducerea la îndeplinire a aceluiași buget în unul din anii următori, când prețurile vor deveni $p_1^1, p_2^1, \ldots, p_n^1$, va necesita suma:

$$S_1 = q_1p_1^1 + q_2p_2^1 + \ldots + q_np_n^1 = \sum qp^1.$$

Fiind vorba despre măsura variației relative, vom avea, raportând pe S_1 la S_0, un indice de variație al sumei stabilite prin bugetul considerat, de la anul de bază la anul considerat:

$$I_{1/0} = \frac{\sum qp^1}{\sum qp^0}. \qquad (157)$$

În alte cazuri se caută să se urmărească variația lui S cauzată nu numai de mișcarea prețurilor, ci și de o evoluție a însești compoziției bugetului.

În acest caz:

$$S_0 = q_1^1p_1^0 + q_2^1p_2^0 + \ldots + q_n^1p_n^0 = \sum q^1p^0$$

$$S_1 = q_1^1p_1^1 + q_2^1p_2^1 + \ldots + q_n^1p_n^1 = \sum q^1p^1$$

și deci:

$$I_{1/0} = \frac{\sum q^1p^1}{\sum q^1p^0}. \qquad (157\text{ bis})$$

De fapt, ambele rapoarte (157) și (157 bis) se exprimă în procente. Amănuntele de calcul pentru aplicarea primei formule reies din tabelul 156. Indicele corespunzător indică pentru București variația costului alimentelor necesare unei persoane din clasa mijlocie. Observăm aici că ponderile q_1, q_2, ..., q_n, deși exprimate în cantități, nu trebuie privite ca niște valori absolute: orice serie de ponderi proporționale cu cele din tabel va conduce la aceleași rezultate. Prin urmare, aceste valori nu au decât o semnificație relativă.

TABELUL 156

Calculul indicelui variației costului alimentelor pentru București

Articole	Cantitatea	Prețul pe unitate			Costul		
		1930	1931	1932	1930	1931	1932
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Carne de vacă	1,500 kg	36,85	29,50	21,25	55,28	44,25	31,88
Carne de porc	0,375 kg	55,90	45,75	31,15	20,96	17,16	11,68
Găină	1/3 buc.	66,65	57,10	52,85	22,22	19,03	17,62
Mezeluri	0,100 kg	82,10	72,10	56,50	8,21	7,21	5,65
Untură	0,250 kg	66,35	42,90	29,75	16,59	10,73	7,44
Unt proaspăt	0,150 kg	149,00	101,00	92,50	22,35	15,15	13,88

Scanned with OKEN Scanner

(continuare tab. 156)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Ouă	5 buc.	2,35	1,98	2,07	11,75	9,90	10,35
Lapte	1,750 kg	12,65	11,35	9,35	22,14	19,86	16,36
Brânză albă	0,150 kg	57,10	46,25	38,50	8,57	6,94	5,78
Untdelemn	0,150 kg	40,00	37,35	28,65	6,00	5,60	4,30
Zahăr tos	0,500 kg	35,50	37,85	29,65	17,75	18,93	14,83
Făină albă	0,300 kg	11,10	9,65	8,85	3,33	2,90	2,66
Pâine albă	3,000 kg	9,90	7,15	8,00	29,70	21,45	24,00
Mălai	0,200 kg	5,15	4,65	3,50	1,03	0,93	0,70
Cartofi	1,000 kg	3,35	3,60	3,70	3,35	3,60	3,70
Ceapă	0,500 kg	2,90	3,40	3,75	1,45	1,70	1,88
Fasole	0,200 kg	11,15	8,40	6,35	2,23	1,68	1,27
Orez	0,100 kg	27,00	23,90	21,75	2,70	2,39	2,18
Cafea	0,075 kg	155,00	130,00	109,60	11,63	9,75	8,22
					267,24	**219,16**	**184,38**

Luând ca termen de comparație anul 1930, care poartă din această cauză numele de *an de bază*, avem:

	1931	1932
indicele variației costului alimentelor	$\frac{219,16}{267,24} = 82,0$	$\frac{184,38}{267,24} = 69,0$

Valoarea indicelui pentru anul de bază este evident egală cu 100.

198. *Indici monetari.* Indicele monetar prezintă o structură mai delicată. Introducerea acestui instrument statistic s-a născut din relativitatea felului cum noi putem privi noțiunea de preț. Prețul unor mărfuri se poate exprima fie prin numărul de unități monetare necesare achiziției unei unități din acea marfă, fie prin numărul de unități-marfă ce pot fi obținute în schimbul unei unități-monedă. Dacă marfa considerată se *ieftinește*, prețul ei exprimat în primul din modurile menționate *descrește*, prețul exprimat în unități-marfă pe unitate-monedă *crește*. Lucrul privit din acest din urmă punct de vedere ne îndreptățește să spunem că moneda s-a scumpit. O *ieftinire* a monedei ar determina deci o *scumpire* a mărfurilor și viceversa.

Cantitatea de monedă destinată cumpărării unei anumite mărfi nu este singurul factor care determină prețul acesteia. Lăsând deoparte faptul că această cantitate însăși poate varia după anotimp, modă sau importanța mărfii considerate, prețul mai depinde și de cantitatea de marfă ce se poate găsi de vânzare la un moment dat. Așadar, chiar dacă am presupune că moneda s-ar ieftini în raportul 1 : 2, nu toate prețurile se vor scumpi în raportul invers 2 : 1, ci fiecare va avea o variație proprie care va depinde de cauze individuale ale târgului acestei singure mărfi.

Scanned with OKEN Scanner

Indicele monetar este, prin definiţie, instrumentul care va măsura variaţia ansamblului de preţuri datorată numai variaţiei valorii monedei.

Construcţia acestui indice constituie o problemă dintre cele mai controversate. Diferite puncte de vedere au fost adoptate pentru rezolvarea ei.

Vom începe cu punctul de vedere care a fost utilizat de la început.

Primii indici au fost calculaţi fără a se căuta să se stabilească printr-o definiţie prealabilă ceea ce se înţelegea să se măsoare prin ei. Majoritatea au fost construiţi într-un mod arbitrar deci, deşi cu oarecare circumspecţie.

S-au întrebuinţat, în acest scop, aproape numai după criterii de afinităţi personale faţă de acestea, toată scara de valori medii mai cunoscute.

Primele încercări de construcţie a indicilor de preţ au pornit de la ideea că aceştia trebuie să exprime mişcarea de ansamblu a preţurilor şi deci trebuie să recurgă la una din valorile medii. Mai târziu, faptul că în calculul acestei valori nu toate elementele trebuie să ocupe aceeaşi importanţă a condus la ideea de ponderare.

Astăzi, încercările de a pune pe baze ştiinţifice construcţia indicilor monetari se sprijină pe trei puncte de vedere diferite:

1. Formal;
2. Monetar (prin utilizarea pseudoecuaţiei teoriei cantitative a monedei);
3. Statistic.

Le vom examina pe rând.

199. *Determinarea formulei indicelui de preţ pe bază de criterii formale.* Profesorul IRVING FISHER şi L. MARCH au fost cei dintâi care au întreprins o analiză în această direcţie[1]. Punctul lor de vedere a fost pur formal. Este clar că însuşi scopul pentru care este construit un indice de preţ ne conduce la anumite condiţii, independente de elementele care intră în compoziţia indicelui şi în afară de orice consideraţie în legătură directă cu acesta, pe care formula lui trebuie să le îndeplinească.

Iată condiţiile ce constituie criteriul profesorului L. MARCH:

1. Indicele trebuie să fie sensibil, adică să reliefeze cele mai mici variaţii ale preţurilor[2].
2. Indicele trebuie să fie independent de unitatea de măsură, altfel obiceiuri diferite de a concretiza preţurile mărfurilor ar avea o influenţă nemotivată asupra valorii indicilor.
3. Când toate preţurile cresc sau descresc în acelaşi raport, indicele trebuie să prezinte aceeaşi variaţie proporţională.
4. Indicele trebuie să fie independent faţă de anul ales ca bază, criteriu care poartă numele de *condiţie de circularitate*.

---

[1] *I. Fisher,* ***The Making of Index Numbers,*** *Boston, Houghton Mifflin, 1922; L. March,* ***Les modes de mesure du mouvement général des prix,*** *Metron, vol. 1, 3, p. 57.*

[2] *Profesorul F. Y. Edgeworth contestă acest criteriu, motivând că, dimpotrivă, un indice de preţ nu trebuie să fie influenţat prea mult de variaţiile capricioase ale unui număr restrâns de articole.*

Scanned with OKEN Scanner

5. Formula trebuie să fie cât se poate de simplă pentru a fi uşor interpretată în raport cu realitatea.

Condiţiile impuse de prof. FISHER[1] sunt ceva mai complicate, deoarece el leagă chestiunea indicelui de preţ de aceea a indicelui activităţii schimbului. Nu vom menţiona deocamdată decât condiţiile relative la indicele de preţ:

1. Condiţia de variaţie proporţională (3, March).
2. Un indice de preţ nu trebuie să devină nedeterminat, nul sau infinit[2].
3. Un indice nu trebuie să varieze dacă se adaugă sau se înlătură din compoziţia lui un articol al cărui preţ relativ este egal cu indicele considerat.
4. Condiţia de circularitate (4, March)[3].
5. Condiţia relativă la schimbarea unităţii de măsură (2, March).

Dacă pe baza unui grup de criterii formale - spun un grup, deoarece numărul şi felul lor nu ne este impus de nimic - s-ar fi putut ajunge la determinarea unei singure formule pentru indicii de preţ, poate că punctul de vedere formal n-ar fi fost aşa de viu atacat. Aşa, de pildă, FISHER ajunge la concluzia că nici una din cele 134 de formule pe care le examinează, şi este greu de conceput că ar mai exista şi altele demne de luat în seamă, nu îndeplineşte absolut toate criteriile impuse de el. N-ar urma de aici că însăşi noţiunea de nivel al preţurilor este şubredă? Chiar faptul că aceste criterii diferă de la un autor la altul constituie o dovadă a imperfecţiunii procedeului de selecţionare. Totuşi nu trebuie să fim prea exigenţi. Cercetările lui MARCH şi FISHER au netezit în bună parte terenul şi au pus capăt alegerii arbitrare a formulei indicelui de preţ.

Vom trece acum să examinăm, în lumina criteriilor de mai sus, câteva din valorile tipice mai răspândite. Să considerăm indicele calculat pe baza *mediei aritmetice a preţurilor relative*, adică a raporturilor dintre preţurile din epoca curentă şi cele ale anului de bază.

Notând preţul relativ prin:

---

[1] *I. Fisher,* ***The Purchasing Power of Money****, New York, Macmillan, 1911.*

[2] *Condiţia aceasta mi se pare de prisos, deoarece nu văd cum cazurile menţionate s-ar putea prezenta efectiv în practică.*

[3] *În lucrarea sa ulterioară,* ***The Making of Index Numbers****, Fisher revine asupra acestui criteriu, opinând că el nu trebuie îndeplinit decât sub forma mai restrânsă a reversibilităţii în timp sau spaţiu. Ceva mai mult, Fisher susţine în această din urmă lucrare că* « the circular test ought not to be fulfiled », *bizuindu-se pe faptul că ar fi absurd să comparăm nivelul preţurilor din Norvegia şi Egipt printr-o comparaţie indirectă a situaţiei din aceste două ţări cu aceea din Georgia. Înclin să cred că problema nu se pune astfel. Este ştiut că două situaţii economice prea diferite nu se pot compara - fapt care a dus la cunoscutele comentarii asupra indicilor istorici - însă în cadrul de aplicabilitate al indicilor de preţ, condiţia de circularitate trebuie să fie printre cele dintâi pe care un indice trebuie să le satisfacă. Din cauza absurdităţii la care ar da naştere într-un caz în care indicii nici nu sunt aplicabili, nu trebuie să se renunţe la o aşa de esenţială proprietate a acestora.*

Scanned with OKEN Scanner

$$\pi^1 = \frac{p^1}{p^0}$$

vom avea:

$$I_{1/0} = M_s = \frac{1}{n}\sum\left(\frac{p^1}{p^0}\right) = \frac{1}{n}\sum\pi^1.$$

Constatăm în primul rând că această formulă este independentă de unitatea de măsură, căci, dacă presupunem că în loc de unitatea u se adoptă unitatea $u' = \frac{u}{\tau}$, atunci noile prețuri vor fi:

$$r^1 = p^1\tau \qquad r^0 = p^0\tau$$

și deci prețul relativ rămâne neschimbat:

$$\frac{r^1}{r^0} = \pi^1.$$

Condiția de proporționalitate este și ea îndeplinită; dacă toate prețurile sunt înmulțite cu același factor α, și indicele va fi modificat la fel:

$$\frac{1}{n}\sum\pi^1\alpha = \alpha\cdot\frac{1}{n}\sum\pi^1.$$

Media M_s satisface și condiția 3, Fisher; dacă $\pi_n^1 = M_s$, atunci:

$$M_s = \frac{n-1}{n}\cdot\frac{\sum_1^{n-1}\pi^1}{n-1} + \frac{M_s}{n}$$

și deci:

$$\frac{\sum_1^{n-1}\pi^1}{n-1} = M_s.$$

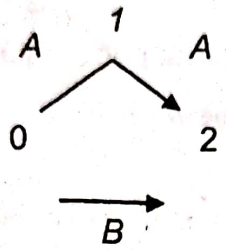
Această valoare tipică nu îndeplinește însă condiția de circularitate.

Înțelegem prin această condiție proprietatea următoare: presupunem că am calculat indicii anului 1 și 2 în raport cu anul de bază, precum și pe acela al

anului 2 în raport cu anul 1. Dacă formula întrebuințată ne conduce la rezultate care verifică egalitatea:

$$I_{1/0} \times I_{2/1} = I_{2/0}$$

vom spune că aceasta îndeplinește condiția de circularitate. Termenul întrebuințat este destul de sugestiv, exprimând faptul că putem obține valoarea indicelui $I_{2/0}$ pe o cale ocolită, mergând adică de la anul de bază la anul 1 și de aici la anul 2. Cu alte cuvinte, drumurile *A* și *B* din schema:



sunt echivalente.

Tot o formă particulară a circularității este exprimată prin formula:

$$I_{1/0} = \frac{1}{I_{0/1}}$$

condiție care poartă numele de condiție de reversibilitate.

Această condiție exprimă faptul următor:

Dacă, luându-se ca bază anul 0, indicele corespunzător anului 1 ne arată o creștere proporțională egală cu *c* în nivelul general al prețurilor, atunci, luându-se ca bază anul 1, indicii astfel obținuți vor trebui să ne arate aceeași creștere proporțională de la un an la altul.

În cazul mediei aritmetice simple:

$$I_{1/0} = \frac{1}{n}\sum \pi^1 \qquad I_{0/1} = \frac{1}{n}\sum \frac{1}{\pi^1}$$

ori:

$$\sum \pi^1 \times \sum \frac{1}{\pi^1} - \left(\sum \frac{\sqrt{\pi^1}}{\sqrt{\pi^1}}\right)^2 \geq 0$$

$$\sum \pi^1 \times \sum \frac{1}{\pi^1} \geq n^2 \ ^1$$

[1] *După formula lui Lagrange.*

deci:

$$I_{1/0} \times I_{0/1} \geq 1$$

Formulele care nu sunt nici reversibile au fost numite « tendenţioase »[1].

Să ilustrăm printr-un exemplu numeric neajunsul unei formule tendenţioase. Să considerăm tabelul următor:

Articole	Preţuri		Preţuri relative	
	Anul 0 p_0	Anul 1 p_1	$\frac{p^1}{p^0}$	$\frac{p^0}{p^1}$
A	25	20	80	125
B	8	14	175	57
C	12	17	142	71
D	19	10	53	190
			450	**443**

Vom avea:

$$I_{1/0} = 112 \qquad I_{0/1} = 111$$

ceea ce înseamnă că, în orice sens am considera mişcarea preţurilor, ea se traduce întotdeauna printr-o majorare şi, ceea ce este şi mai curios, această majorare este aceeaşi.

De aici se deduce că într-o serie de indici calculaţi pe baza mediei aritmetice a preţurilor relative, dacă la doi din anii consideraţi le corespund indicii 160 şi 200, de pildă, nu înseamnă că nivelul general al preţurilor a suferit o creştere de 25% de la un an la altul.

Media geometrică simplă conduce la formula:

$$I_{1/0} = \sqrt[n]{\pi_1^1 \pi_2^1 \ldots \pi_n^1}$$

sau:

$$\log I_{1/0} = \frac{1}{n} \sum \log \pi^1$$

În ceea ce priveşte sensibilitatea (condiţia 1, March) media geometrică se bucură de o proprietate care face din ea o formulă utilă în special pentru măsura nivelului preţurilor. Dacă unul din preţurile relative π^1 suferă o creştere $\Delta\pi^1$, atunci creşterea corespunzătoare indicelui va fi:

---

[1] *În engleză,* biased.

$$\Delta I_{1/0} = \frac{1}{n} \cdot \frac{I_{1/0}}{\pi^1} \cdot \Delta\pi^1$$

ceea ce ne arată că variaţia lui $I_{1/0}$ va fi cu atât mai mare cu cât raportul dintre $I_{1/0}$ şi π^1 va fi mai ridicat şi viceversa, adică variaţiile preţurilor articolelor ieftine (al căror preţ relativ se găseşte sub nivelul general al preţurilor) vor influenţa mai mult indicele decât variaţia articolelor scumpe.

Este uşor să ne asigurăm că media geometrică îndeplineşte condiţia de proporţionalitate, condiţia 3, Fisher, şi că este independentă de unitatea de măsură.

Media geometrică verifică condiţia de circularitate; într-adevăr:

$$I_{1/0} \times I_{2/1} = \sqrt[n]{\frac{p_1^1}{p_1^0} \cdot \frac{p_2^1}{p_2^0} \cdots \frac{p_n^1}{p_n^0} \times \frac{p_1^2}{p_1^1} \cdot \frac{p_2^2}{p_2^1} \cdots \frac{p_n^2}{p_n^1}} = \sqrt[n]{\frac{p_1^2}{p_1^0} \cdot \frac{p_2^2}{p_2^0} \cdots \frac{p_n^2}{p_n^0}} = I_{2/0}.$$

Acest fapt ne permite să comparăm între ei indicii corespunzând la ani diferiţi, oricare ar fi poziţia lor în raport cu anul de bază.

O altă formulă întrebuinţată pentru calculul indicilor de preţ este *agregatul simplu* [1]:

$$I_{1/0} = \frac{\sum p^1}{\sum p^0},$$

dar care are marele neajuns de a nu fi independent faţă de unitatea de măsură. Or, cum unitatea de măsură a unui articol nu ne este prin nimic impusă din afară, urmează că variaţiile acestuia vor fi influenţate de alegerea pe care o facem noi asupra unităţilor de măsură, ceea ce este arbitrar. Dacă printre preţurile componente socotim, de pildă, şi pe acela al tonei de radium, atunci indicele nu va acuza decât variaţiile datorate acestui singur articol.

Acest fapt face ca indicele astfel calculat să fie printre cei mai defectuoşi. Cu toate acestea, indicele publicat de revista americană *Bradstreet*, deşi construit după acest tip, este unul din indicii cei mai buni din Statele Unite. Explicaţia se găseşte într-o judicioasă alegere a articolelor din compoziţia lui, precum şi a întrebuinţării unor unităţi de măsură adecvate - o probă a faptului că formula de calcul trece, în mod practic, pe planul al doilea de importanţă atunci când este vorba să construim un indice.

S-a examinat şi posibilitatea reprezentării nivelului preţurilor prin valoarea mediană a preţurilor relative π_1. Prof. F. Y. EDGEWORTH în special recomandă acest procedeu care, după cum ştim, conduce la o valoare din cele mai stabile.

---

[1] *Acesta este obţinut prin inversarea operaţiilor care au condus la media aritmetică a raporturilor dintre preţuri, adică se calculează întâi mediile preţurilor absolute şi apoi se face raportul dintre ele.*

Scanned with OKEN Scanner

Faptul că mediana prezintă totuși o abatere tip datorată hazardului mai mare decât aceea a mediei aritmetice (§ 103) constituie un argument contra adoptării acestui tip de indice în cazul când avem la dispoziție numai prețurile unui număr restrâns de articole.

Mediana conduce la un indice reversibil, nu însă circular. Într-adevăr, ordinea elementelor păstrându-se prin inversarea valorilor, dacă $I_{1/0} = \pi^1 = \frac{p^1}{p^0}$, vom avea $I_{0/1} = \frac{1}{\pi^1} = \frac{p^0}{p^1}$. Punând:

$$I_{1/0} = \frac{p_i^1}{p_i^0} \text{ și } I_{2/1} = \frac{p_k^2}{p_k^1}$$

chiar dacă $i = k$, nimic nu ne obligă să avem și $I_{0/2} = \frac{p_i^0}{p_i^2}$.

200. *Formule ponderate*. Formulele ponderate se împart în două categorii: cu ponderi constante și cu ponderi variabile. Dintre formulele cu ponderi constante, cea mai răspândită este aceea a *agregatului ponderat prin cantități*:

$$A_p = \frac{\sum qp^1}{\sum qp^0}. \tag{158}$$

Cantitățile q rămân în acest caz constante de-a lungul seriei și pot fi determinate după cele mai variate criterii, care depind în primul rând de întrebuințarea care i se va da indicelui.

Agregatul ponderat se mai poate pune sub forma:

$$M_p = \frac{\sum qp^0 \cdot \pi^1}{\sum qp^0} \tag{158 bis}$$

care este o medie aritmetică ponderată a prețurilor relative π^1. Singura deosebire dintre formulele precedente constă în faptul că, în timp ce pentru agregat ponderea apare ca o *cantitate q*, în media aritmetică ponderea este o *valoare qp*. Acest fapt are importanța lui, deoarece mulți statisticieni au fost tentați, din cauza diferenței menționate, să considere cele două formule complet deosebite una de alta.

Indicele calculat pe baza formulei (158) a fost recomandat încă din 1864 de către LASPEYRES. Este unul din procedeele cele mai răspândite pentru

Scanned with OKEN Scanner

calculul indicilor. În afară de faptul că îndeplinește toate condițiile formale, această formulă are avantajul de a corespunde și punctului de vedere monetar[1]. O interpretare plastică a formulei lui Laspeyres a fost dată de către J. RUEFF[2].

În mod teoretic și în cazul unui târg ideal de publicitate deplină a cererilor și ofertelor, prețul unei mărfi la un moment dat se stabilește după legea exprimată prin formula:

$$p = \frac{s}{q} \tag{159}$$

unde s este suma totală pe care toți indivizii care iau parte la târg sunt dispuși să o ofere pentru marfa considerată, iar q este cantitatea disponibilă la momentul considerat.

Dacă adoptăm acum ca unitate de măsură fiecare marfă în parte, cantitatea ce se poate obține în schimbul unei unități monedă[3], prețurile fiind cele din anul de bază, vom avea pentru (158):

$$I_{1/0} = \frac{\sum qp^1}{\sum q}$$

și cum $\sum qp^1$ = suma tuturor bugetelor, iar $\sum q$ = cantitatea totală de mărfuri ce au făcut obiectul unei tranzacții, vom avea:

$$I_{1/0} = \frac{S}{Q}. \tag{159 bis}$$

Formula indicelui monetar pusă sub forma (159 bis) ne arată că indicele nivelului general al prețurilor nu se modifică decât dacă variază suma totală de care indivizii colectivității considerate dispun spre a face *toate* cumpărăturile necesare sau dacă variază cantitatea totală a bunurilor disponibile pe piață. Această analogie între stabilirea prețului unui singur articol și aceea a nivelului general al prețurilor vine să mărească și mai mult valabilitatea formulei lui Laspeyres.

Din cauză însă că prețurile stau în strânsă legătură cu cantitățile și că factorul monetar le influențează pe amândouă, s-au propus și indici în care ponderile să fie variabile, adoptând pentru fiecare an în parte cantitățile respective.

---

[1] *N. St. Georgescu,* ***Tehnica numerelor indice pentru nivelul general al prețurilor,*** *Bul. statistic al României, nr. 4, p. 116.*

[2] *J. Rueff,* ***Théorie des phénomènes monétaires****, Paris, Payot, 1927.*

[3] *Dollar's worth.*

Ne vom opri asupra câtorva din aceste tipuri, cele pe care Fisher le consideră cele mai bune din punct de vedere formal.

Avem în primul rând agregatul ponderat prin cantităţile corespunzătoare fiecărui an în parte:

$$I_{1/0} = \frac{\sum q^1 p^1}{\sum q^1 p^0}$$

expresie care prezintă o mare analogie cu agregatul ponderat prin coeficienţi constanţi. Nu îndeplineşte însă condiţia de circularitate.

Fisher are, de asemenea, o deosebită încredere în indicii definiţi prin formulele:

$$\frac{\sum (q^0 + q^1) p^1}{\sum (q^0 + q^1) p^0} \quad , \quad \frac{\sum \sqrt{q^0 q^1} \cdot p^1}{\sum \sqrt{q^0 q^1} \cdot p^0}$$

care tind să ţină seama şi de importanţa schimbului din anul de bază şi de cea din anul corespunzător indicelui.

În fine, formula *ideală* a lui Fisher este construită în acelaşi scop:

$$I_{1/0} = \sqrt{\frac{\sum q^0 p^1}{\sum q^0 p^0} \cdot \frac{\sum q^1 p^1}{\sum q^1 p^0}}.$$

Ultimele trei formule îndeplinesc toate condiţiile Fisher, cu excepţia criteriului de circularitate, pe care nu-l satisfac decât sub forma mai simplă a reversibilităţii.

Cu toată recomandarea călduroasă pe care profesorul Irving Fisher o face acestor metode de calcul personal, utilizează pentru indicele pe care îl publică săptămânal formula lui Laspeyres.

201. *Analiza statistică a indicilor de preţ.* Mai există încă un mod de a privi problema indicelui de preţuri. Este vorba despre teoria statistică a acestui instrument de urmărire a variaţiilor preţurilor. Acest din urmă punct de vedere a fost prea puţin studiat, deşi unele din chestiunile ce compun această problemă au fost în parte atinse accidental.

Ca orice problemă rezolvată prin metodele statistice moderne, acest studiu statistic al preţurilor trebuie să înceapă printr-o examinare a distribuţiei acestora.

Studii în această direcţie au fost făcute de W. MITCHEL, N. CRUMP, L.MARCH şi M. OLIVIER[1].

[1] *M. Olivier,* ***Les nombres indices de la variation des prix****, Paris, Giard, 1927.*

Scanned with OKEN Scanner

Concluziile la care s-a ajuns sunt următoarele:

1. Distribuţia preţurilor relative este departe de a fi normală. Curba corespunzătoare este negativ asimetrică.
2. Distribuţia logaritmilor preţurilor relative se face după o curbă simetrică leptokurtică. Acest ultim rezultat ne îndreptăţeşte să adoptăm ca valoare tipică a unui ansamblu de preţuri media aritmetică a logaritmilor lor, care ne conduce de fapt la media geometrică a preţurilor.

Aşadar, analiza statistică a distribuţiei preţurilor impune un indice calculat pe baza mediei geometrice. De fapt, rezultatele menţionate ne permit stabilirea în cazul de faţă a unei egalităţi între mediana preţurilor şi media geometrică a lor; dacă egalitatea lor nu are loc în practică este din cauza fluctuaţiilor hazardului. Este totuşi preferabil să utilizăm pentru determinarea indicelui media geometrică, mediana fiind într-adevăr prea puţin sensibilă.

202. *Indicii de producţie*. La fel cum indicele de preţ pune în evidenţă mişcarea de ansamblu a preţurilor, indicele de producţie va reliefa variaţia globală a producţiei din diferitele ramuri de activitate.

Indicii de producţie se obţin imediat din formulele întrebuinţate pentru indicii de preţ, schimbând preţurile şi cantităţile între ele. Astfel, indicele asociat indicelui de preţ calculat pe baza mediei aritmetice este dat de formula:

$$Q_{1/0} = \frac{1}{n}\sum\left(\frac{q^1}{q^0}\right).$$

Indicele asociat formulei lui Laspeyres este:

$$Q_{1/0} = \frac{\sum pq^1}{\sum pq^0},$$

iar cel asociat formulei ideale a lui Fisher:

$$Q_{1/0} = \sqrt{\frac{\sum p^0q^1}{\sum p^0q^0} \cdot \frac{\sum p^1q^1}{\sum p^1q^0}}.$$

De fapt, din punct de vedere formal, orice condiţie impusă indicelui de preţ trebuie să fie verificată şi de $Q_{1/0}$.

Printre condiţiile formale impuse de Fisher, în afară de cele menţionate la timp, mai există una care priveşte deopotrivă indicii asociaţi de preţ şi producţie şi care se traduce prin relaţia:

$$I_{1/0} \times Q_{1/0} = V_{1/0}\ ^1$$

unde $V_{1/0}$ este raportul dintre valorile mărfurilor schimbate în cei doi ani consideraţi:

---

[1] *O traducere formală a faptului că valoarea este egală cu produsul dintre preţ şi cantitate.*

$$V_{1/0} = \frac{\sum p^1 q^1}{\sum p^0 q^0}.$$

Evident că numai indicii ponderați prin cantități variabile pot îndeplini acest din urmă criteriu. În particular, el este verificat de către formula ideală.

În ceea ce privește determinarea cantităților din compoziția lui, vom avea de-a face cu un indice de producție, dacă utilizăm cantitățile produse, sau cu un indice al activității schimbului, dacă ne bazăm calculele pe cantitățile vândute.

203. *Considerații asupra alegerii ponderilor și a anului de bază.* În ceea ce privește ponderile, acestea nu influențează prea mult variația unui indice, cu condiția expresă ca stabilirea lor să nu se facă în mod arbitrar, ci după un criteriu logic și unitar[1].

Nu putem spune însă același lucru despre anul de bază, atunci când formula utilizată pentru calculul indicelui nu îndeplinește condiția de circularitate. S-a afirmat categoric chiar, că schimbarea anului de bază poate influența numai nivelul indicilor, dar în nici un caz sensul evoluției lor. Exemplul numeric care urmează ne va dovedi netemeinicia acestei afirmații.

Articole	Prețuri		
	Anul 0	Anul 1	Anul 2
A	12	72	80
B	15	60	75
C	30	30	10

Luând ca bază primul an, indicii calculați pe baza mediei aritmetice vor avea următoarele valori:

$$I_{0/0} = 100 \qquad I_{1/0} = 367 \qquad I_{2/0} = 400.$$

Aceleași date ne conduc însă la indicii:

$$I_{0/1} = 47 \qquad I_{1/1} = 100 \qquad I_{2/1} = 90$$

care nu mai prezintă același sens de variație ca cei dintâi.

Pentru a evita ca acest defect al indicilor necirculari (căci pentru cei circulari problema nici nu se pune) să nu conducă la diferențe prea pronunțate, trebuie ca alegerea anului de bază să fie făcută cu grijă și de preferință într-unul din anii seriei care poate fi considerat ca normal din punctul de vedere al fenomenului ale cărui variații le urmărim. O atenție specială trebuie acordată fenomenelor cu neregularități de variație simțitoare, cum este, de pildă, cazul indicilor cursurilor valorilor mobiliare.

[1] *A. Bowley, **op. cit.**, p. 87.*

Scanned with OKEN Scanner

Apendice

204. *Variabilă întâmplătoare*. Orice distribuție heterogradă este un exemplu de variabilă întâmplătoare. Presupunând că avem de-a face cu o variabilă discontinuă, variabila putând lua numai valorile (x_1 , x_2, ..., x_s), respectiv cu frecvențele (f_1 , f_2, ..., f_s), aceasta va fi complet definită prin elementul:

$$X\begin{pmatrix} x_1, x_2, \ldots\ldots, & x_s \\ f_1, f_2, & \ldots\ldots, \quad f_s \end{pmatrix}.$$

Problema se complică însă în cazul unei variabile continue. În acest din urmă caz, vom determina pentru fiecare valoare x a variabilei întâmplătoare X, frecvența cazurilor pentru care $X < x$. Obținem astfel o funcție $N(x)$ care se identifică de fapt cu noțiunea de frecvență cumulată și pe care o vom denumi *funcția frecvențelor cumulate*[1]. Reprezentarea grafică a acestei funcții dă naștere la ogiva corespunzătoare distribuției considerate. Faptul că funcția $N(x)$ poate fi determinată indiferent dacă variabila este continuă sau nu face din aceasta un element extrem de util pentru cercetări. Singura deosebire între un caz și altul este alura funcției $N(x)$, deosebire pe care am accentuat-o la vreme (§ 52).

Funcția $N(x)$ este evident o funcție nedescrescătoare, cuprinsă în cazul frecvențelor relative între 0 și 1, iar dacă variabila nu poate lua valori infinite cu frecvențe finite - caz în care se apropie cel mai mult de distribuțiile întâlnite în practică - vom avea și:

$$N(-\infty) = 0 \qquad N(+\infty) = 1.$$

Frecvența cazurilor pentru care $a \leq X < b$ va fi evident egală cu $N(b) - N(a)$; pentru cazul când $x \leq X < x + dx$, vom avea:

$$N(x + dx) - N(x) = \Delta N(x).$$

Dacă $N(x)$ este continuă în punctul x, deducem imediat din relația precedentă că frecvența cazurilor $X = x$ este nulă. Acest fapt constituie o caracteristică a variabilelor întâmplătoare continue. Nu trebuie dedus de aici că

---

[1] *Această funcție este cunoscută în calculul probabilităților sub numele de* funcție de probabilitate totală.

Scanned with OKEN Scanner

$X = x$ constituie o imposibilitate, căci am ajunge astfel la o absurditate evidentă. Numai frecvenţa relativă a cazurilor $X = x$ - care, după cum ştim, este un raport - este limita unui infinit mic şi egală deci cu zero.

Dacă pe deasupra $N(x)$ este o funcţie derivabilă, frecvenţa cazurilor pentru care $x \leq X < x + dx$ va fi egală cu $f(x)dx$, unde $f(x) = N'(x)$. Curba $y = f(x)$ este chiar curba frecvenţelor corespunzătoare distribuţiei considerate. Pe o astfel de curbă, frecvenţa cazurilor $a \leq X < b$ (sau $a < X < b$, căci frecvenţa lui $X = a$ este nulă) va fi reprezentată de suprafaţa cuprinsă între ordonatele $x = a$ şi $x = b$:

$$\int_a^b f(x)dx = N(b) - N(a) .$$

Din consideraţiile precedente se desprinde faptul că funcţia frecvenţelor cumulate defineşte pe deplin orice distribuţie heterogradă.

205. *Medii.* Singure cele două condiţii accentuate de către CAUCHY (§ 64) vin să îngrădească într-o slabă măsură definiţia complet arbitrară a unei valori medii. Aceste condiţii sunt:

1. Dacă x_1 şi x_2 sunt valorile extreme ale unei serii statistice $x_1, x_2, ..., x_s$, atunci media acesteia $M = F(x_1, x_2, ..., x_s)$ va trebui să verifice dubla inegalitate:

$$x_1 \leq M \leq x_s.$$

Ca o consecinţă a acestui fapt, deducem că dacă toate valorile seriei sunt egale între ele, $x = \alpha$, atunci şi $M = \alpha$.

2. Dacă în loc de seria $x_1, x_2, ..., x_s$, avem o nouă serie:

$$\begin{array}{c} x_1, x_1, ..., x_1 \\ x_2, x_2, ..., x_2 \\ ..... \\ x_s, x_s, ..., x_s \end{array}$$

în care fiecare element este repetat de p ori, media acestei din urmă serii va fi egală cu a celei dintâi.

O clasă de valori medii care să cuprindă pe cele mai des întâlnite în practică este definită prin formula:

$$\mu(M) = \frac{\sum \mu(x_i)}{s} \qquad (160)$$

μ fiind o funcţie oarecare; media respectivă se obţine rezolvând această ecuaţie în M. Pentru a evita orice ambiguitate, este necesar să precizăm - în cazul în care funcţia implicită $x = \mu(y)$ are mai multe ramuri - pe care din acestea o vom adopta pentru determinarea lui M.

Scanned with OKEN Scanner

Se vede imediat că, dacă $\mu = x^k$, relaţia (160) ne conduce la media aritmetică de ordinul k; dacă $\mu = \log x$, obţinem media geometrică.

Categoria de medii definite prin relaţia (160) se bucură de o proprietate generală: *admit o teoremă de adiţie* (§ 46).

Să însemnăm prin M' şi M'' mediile seriilor $(x_1, x_2, \ldots, x_k)$, $(x_{k+1}, x_{k+2}, \ldots, x_s)$:

$$\mu(M') = \frac{\sum_1^k \mu(x_i)}{k} \qquad \mu(M'') = \frac{\sum_{k+1}^k \mu(x_i)}{s-k}.$$

Vom avea:

$$\mu(M) = \frac{k\mu(M') + (s-k)\mu(M'')}{s}$$

cu alte cuvinte, cunoscând mediile M' şi M'' a două serii statistice, pentru a obţine media seriei rezultante, vom atribui acestora frecvenţe egale cu numărul de elemente cuprinse în fiecare din ele şi vom calcula media lor pe baza aceleiaşi funcţii μ.

Relaţia (160) ne permite o interpretare grafică. Să considerăm curba $y = \mu(x)$ şi pe această curbă punctele $m_1, m_2, m_3, \ldots, m_s$ de abscise $x_1, x_2, \ldots, x_s$. Centrul de greutate G al acestor puncte va avea coordonatele:

$$\xi = \frac{\sum x_i}{s} \qquad \eta = \frac{\sum \mu(x_i)}{s}.$$

Ducând prin punctul G o paralelă la axa Ox, obţinem, prin intersecţia ei cu curba consideratã, un punct m a cărui abscisă M este dată de ecuaţia $\mu(M) = \eta$. M va fi media seriei $x_1, x_2, x_3, \ldots, x_s$ calculată pe baza funcţiei μ. Servindu-ne de această interpretare grafică a mediilor din clasa examinată, vom stabili o proprietate generală a mediilor aritmetice de diferite ordine. Să considerăm o serie de valori pozitive $x_1, x_2, \ldots, x_s$ şi mediile aritmetice de ordinul 1 şi $k > 1$. Centrele de greutate G şi G' ale punctelor de abscise $x_1, x_2, \ldots, x_s$ situate respectiv pe curbele $y = x$ şi $y = x^k$ vor avea coordonatele:

$$G\left(\xi = \frac{\sum x_i}{s},\ \eta = \xi\right) \qquad G'\left(\xi' = \xi,\ \eta' = \frac{\sum x_i^k}{s}\right).$$

Punctele G şi G' se găsesc, aşadar, pe aceeaşi paralelă la Oy, iar G' în concavitatea curbei $y = x^k$ [1], care în tot intervalul $(0, +\infty)$ este îndreptată spre

[1] *Aceasta este o proprietate binecunoscută a centrului de greutate.*

Scanned with OKEN Scanner

y - i pozitivi. Prin urmare punctul m' de ordonată η' situat pe această din urmă curbă se va găsi la dreapta dreptei $x = \xi$, adică:

$$M_k > M_1 .$$

Să considerăm acum două medii aritmetice de ordinul λ şi ν :M_λ şi M_ν şi să presupunem că $\lambda > \nu$, $\lambda\nu > 0$. Vom avea:

$$M_\lambda^\lambda = \frac{\sum x_i^\lambda}{s} \qquad M_\nu^\nu = \frac{\sum x_i^\nu}{s}$$

iar prin schimbarea de variabilă $x^\nu = z$:

$$M_\lambda^\lambda = \frac{\sum z_i^{\frac{\lambda}{\nu}}}{s} \qquad M_\nu^\nu = \frac{\sum z_i}{s}.$$

Cum însă $\frac{\lambda}{\nu} > 1$, în baza proprietăţii de mai sus, deducem:

$$\left(M_\lambda^\lambda\right)^{\frac{\nu}{\lambda}} > M_\nu^\nu$$

adică:

$$M_\lambda > M_\nu \qquad (161)$$

relaţie care ne permite să comparăm valorile a două medii aritmetice de ordine diferite, însă de acelaşi semn. Să observăm acum că media aritmetică de ordinul zero este nedeterminată. Nedeterminarea este numai aparentă, căci luând logaritmii în ambii termeni ai relaţiei (160), avem:

$$\log M_k = \frac{\log \sum x_i^\nu - \log s}{k}$$

şi aplicând în urmă regula lui Hôpital:

$$\lim_{k=0} \log M_k = \lim_{k=0} \frac{\sum x_i^k \log x_i}{\sum x_i^k} = \frac{\sum \log x_i}{s}$$

deci *media aritmetică de ordinul zero este chiar media geometrică*.

Ţinând seama de acest rezultat, printr-o operaţie de trecere la limită în (161), obţinem $M_l < M_0 < M_k$ unde $l < 0 < k$ şi, prin urmare, $M_l < M_k$ dacă $l < k$.

Scanned with OKEN Scanner

Proprietățile demonstrate în acest paragraf rămân valabile în cazul general, după cum este ușor să ne asigurăm, repetând raționamentul asupra variabilei $X\begin{pmatrix} x_i \\ n_i \end{pmatrix}$, când relația (160) trebuie înlocuită prin:

$$\mu(M) = \frac{\sum n_i \mu(x_i)}{\sum n_i}.$$

Dacă variabila este definită prin funcția frecvențelor cumulate, vom avea de-a face cu limita sumei:

$$\sum \mu(\xi_i) \Delta N(x_i)$$

unde $x_i < \xi_i < x_{i+1}$ și care ne conduce la o integrală Stieltjes. Media calculată pe baza funcției μ va fi obținută din relația:

$$\mu(M) = \int_{-\infty}^{+\infty} \mu(x) dN(x)\ ^{1}.$$

206. *Mediana*. Teoretic, mediana este definită prin ecuația $N(x) = 1/2$. În practică însă, funcția $N(x)$ nu este cunoscută decât pentru anumite valori, limitele intervalelor scării în raport cu care s-a făcut distribuția. Este clar că mediana va fi cuprinsă în intervalul median (x_i, x_{i+1}) pentru care $N_{i-1} < 1/2$, $N_i > 1/2$. Vom determina mediana pe baza supoziției că $N(x)$ variază liniar în acest interval, adică pentru $x_i \le x \le x_{i+1}$ $(= x_i + d)$:

$$N(x) = N_{i-1} + \frac{n_i}{d}(x - x_i).$$

Făcând $N(x) = 1/2$, obținem:

$$x(= Me) = x_i + \frac{1 - 2N_{i-1}}{2n_i} \cdot d$$

sau, în cazul frecvențelor absolute (9):

$$Me = x_i + \frac{N - 2N_{i-1}}{2n_i} \cdot d.$$

---

[1] *Pentru integrala aceasta și altele pe care le vom mai întâlni, putem - în scopul de a nu varia prea mult notațiile - adopta ca limite $-\infty$ și $+\infty$, chiar dacă variabila rămâne într-un interval finit (a, b), căci în acest din urmă caz $dN(x) = 0$ pentru orice valoare exterioară intervalului menționat.*

Scanned with OKEN Scanner

Mediana este strâns legată printr-o proprietate de abaterea medie. Fie e_x abaterea medie în raport cu valoarea x. Funcţia e_x nu este o funcţie analitică peste tot, ci este definită după cum urmează:

$$x \le x_1 \qquad e_x = \frac{\sum_1^s n_r x_r}{N} - x$$

$$x_1 \le x \le x_2 \qquad e_x = \frac{\sum_2^s n_r x_r - n_1 x_1}{N} - \frac{\sum_2^s n_r - n_1}{N} \cdot x$$

.....

$$x_i \le x \le x_{i+1} \qquad e_x = \frac{\sum_{i+1}^s n_r x_r - \sum_1^i n_r x_r}{N} - \frac{\sum_{i+1}^s n_r - \sum_1^i n_r}{N} \cdot x$$

.....

$$x_s \le x \qquad e_x = -\frac{\sum_1^s n_r x_r}{N} + x$$

Prin urmare, e_x este formată dintr-o serie de funcţii liniare ai căror coeficienţi unghiulari *cresc* de la -1 la +1. Într-adevăr, aceştia sunt:

$$-1, -1 + \frac{2N_1}{N}, 1 + \frac{2N_2}{N}, \ldots\ldots\ldots\ldots -1 + \frac{2N_{s-1}}{N}, +1$$

Minimul lui e_x va avea loc pentru punctul în care coeficientul unghiular va schimba de semn sau pentru intervalul în care va fi nul. În primul caz, vom avea $N - 2N_{i-1} < 0$, $N - 2N_i > 0$, iar în al doilea $N - 2N_i = 0$. Recunoaştem aici definiţia intervalului median, care la limită conduce la valoarea mediană. Aşadar, *abaterea medie are cea mai mică valoare atunci când este calculată în raport cu mediana.*

O altă determinare practică a medianei este sugerată de însăşi această proprietate a ei. Reprezentarea grafică a funcţiei e_x dă naştere la o linie poligonală cu concavitatea înspre y - i pozitivi. Prin cele trei puncte mai apropiate de axa Ox vom face să treacă o parabolă de ordinul 2. Este firesc să admitem că abscisa minimului acestei parabole va fi o bună evaluare a medianei teoretice. Calculele, care nu prezintă nici o dificultate, ne arată că în cazul distribuţiilor cu intervale egale, acest punct de vedere ne conduce la aceeaşi formulă (9).

Scanned with OKEN Scanner

207. *Momente*. Momentul de ordinul p al unei distribuţii este dat de una din formulele:

$$m_p = \frac{\sum n_i x_i^p}{\sum e_i} \qquad m_p = \int_{-\infty}^{+\infty} x^p dN(x)$$

după cum variabila este discontinuă sau nu. Observăm că în timp ce prima formulă are întotdeauna un sens, cea de a doua poate, în unele cazuri, să fie pur formală. Cu alte cuvinte, momentele unei distribuţii de variabilă continuă pot să nu existe decât până la un anumit ordin.

Una din proprietăţile cele mai de seamă ale momentelor este constituită din felul cum acestea se comportă faţă de o schimbare de origine. Să punem $y = x - a$:

$$m_p(a) = \frac{\sum n_i y_i^p}{N} = \frac{1}{N}(\sum n_i x_i^p - C_p^1 a \sum n_i x_i^{p-1} + \ldots \pm C_p^k a^k \sum n_i x_i^{p-k} \mp \ldots$$

$$\pm a^p \sum n_i) = m_p - C_p^1 a m_{p-1} + C_p^2 a^2 m_{p-2} - \ldots \pm C_p^k a^k m_{p-k} \mp \ldots \pm a^p$$

iar în cazul când noua origine este chiar media aritmetică a distribuţiei:

$$\mu_p = m_p - C_p^1 m_1 m_{p-1} + C_p^2 m_1^2 m_{p-2} - \ldots \pm C_p^k m_1^k m_{p-k} \mp \ldots$$

$$\pm C_p^{p-2} m_1^{p-2} m_2 \mp \left(C_p^{p-1} - 1\right) m_1^p$$

formule de a căror importanţă practică am avut ocazia să ne convingem în capitolele precedente.

208. *Corecţia Sheppard*. În cazul unei distribuţii de variabilă continuă momentele sunt date de relaţia

$$m_p = \int_{-\infty}^{+\infty} x^p f(x) dx$$

în timp ce, în practică, noi nu putem determina decât momentele calculate

$$m'_p = \sum f_i x_i^p$$

unde f_i este frecvenţa valorilor cuprinse în intervalul $x_i - \frac{d}{2}$, $x_i + \frac{d}{2}$, adică:

$$f_i = \int_{x_i - \frac{d}{2}}^{x_i + \frac{d}{2}} f(x) dx = \int_{-\frac{d}{2}}^{+\frac{d}{2}} f(x_i + y) dy .$$

Scanned with OKEN Scanner

Prin urmare:

$$m'_p = \sum x_i^p \int_{-\frac{d}{2}}^{+\frac{d}{2}} f(x_i + y)dy$$

valoare care evident diferă simţitor de m_p. Dezvoltând în serie Taylor şi integrând:

$$m'_p = \sum x_i^p \int_{-\frac{d}{2}}^{+\frac{d}{2}} \left[f(x_i) + \frac{y}{1!} Df(x_i) + \ldots + \frac{y^k}{k!} D^k f(x_i) + \ldots \right] dy = \tag{162}$$

$$= \sum_0^\infty \frac{d^{2k+1}}{(2k-1)!} \cdot \frac{1}{2^{2k}} \sum x_i^p D^{2k} f(x_i) .$$

Mai departe, Sheppard presupune că $f(x)$ şi derivatele sale se anulează la extremităţile intervalului de variaţie reală a lui x, iar dacă acest interval este $(-\infty, +\infty)$, atunci pentru aceste valori $x^k D^j f(x) = 0$.

În baza ipotezei făcute vom avea:

$$\left| D^{r-1} \left(x^p D^{2k} f \right) \right|_a^b = 0 .$$

Cu ajutorul formulei lui Euler-Maclaurin:

$$d \sum F(x_i) = \int_a^b F(x)dx + \sum_1^m a_r \left| D^{r-1} F(x) \right|_a^b + R_m$$

obţinem, neglijând restul R_m:

$$d \sum x_i^p D^{2k} f(x_i) = \int_a^b x^p D^{2k} f(x)dx .$$

Vom avea, de asemenea:

$$\int_a^b x^p D^{2k} f(x)dx = \left| x^p D^{2k-1} f(x) \right|_a^b - p \int_a^b x^{p-1} D^{2k-1} f(x)dx =$$

$$= -p \int_a^b x^{p-1} D^{2k-1} f(x)dx = p(p-1)\ldots(p-2k+1) \int_a^b x^{p-2k} f(x)dx$$

Scanned with OKEN Scanner

dacă $p \geq 2k$, şi:

$$\int_a^b x^p D^{2k} f(x)dx = 0$$

pentru $p < 2k$.

Ţinând seama de aceste rezultate, relaţia (162) devine:

$$m'_p = \sum_{k=0} \frac{p(p-1)..(p-2k+1)}{(2k+1)!}\left(\frac{d}{2}\right)^{2k} m_{p-2k}$$

adică:

$$m'_0 = m_0 \qquad m'_1 = m_1 \qquad m'_2 = m_2 + \frac{d^2}{12}m_0$$

$$m'_3 = m_3 + \frac{d_2}{4}m_1 \qquad m'_4 = m_4 + \frac{d^2}{2}m_2 + \frac{d^4}{80}m_0$$

de unde deducem, în cazul când originea este egală cu media aritmetică (41):

$$\mu_2 = \nu_2 - \frac{d^2}{12} \qquad \mu_3 = \nu_3 \qquad \mu_4 = \nu_4 - \frac{d^2}{2}\nu_2 + \frac{7d^4}{240}.$$

209. *Funcţie caracteristică. Semiinvarianţi.* Presupunând că suntem în cazul unei distribuţii pentru care se pot determina toate momentele, funcţia generatoare a şirului:

$$m_0, m_1, m_2, ..., m_p, ...$$

adică:

$$\varphi(t) = 1 + m_1\frac{t}{1!} + m_2\frac{t^2}{2!} + ... + m_p\frac{t^p}{p!} + ...$$

poartă numele de *funcţie caracteristică* a distribuţiei considerate. Această funcţie apare în acelaşi timp şi ca media funcţiei e^{xt} calculată în raport cu distribuţia respectivă:

$$\varphi(t) = \int_{-\infty}^{+\infty} e^{xt} dN(x) \qquad (163)$$

sau, în cazul când variabila este discontinuă:

$$\varphi(t) = \sum f_i e^{x_i t}$$

Scanned with OKEN Scanner

Acest fapt ne permite definirea funcției caracteristice indiferent dacă șirul momentelor există în întregime sau nu.

Cunoașterea funcției caracteristice echivalează cu cunoașterea distribuției înseși. Într-adevăr, făcând în (163) schimbarea de variabilă $t = i\tau$, avem:

$$\Phi(\tau) = \int_{-\infty}^{+\infty} e^{ix\tau} f(x)dx .$$

Or, în baza inversiunii integralei lui Fourier:

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} e^{-ix\tau} \Phi(\tau)d\tau \quad [1].$$

Funcția caracteristică definită prin relația (163) nu conduce neapărat la șirul momentelor, căci se poate întâmpla ca aceasta să nu fie desfășurabilă în serie. Legătura formală dintre momente și funcția caracteristică este însă de cea mai mare utilitate în cercetări.

Dacă dezvoltăm în serie log $\varphi(t)$, obținem:

$$\psi(t) = \log \varphi(t) = s_1 \frac{t}{1!} + s_2 \frac{t^2}{2!} + \ldots + s_p \frac{t^p}{p!} + \ldots$$

serie prin intermediul căreia definim noi caracteristici ale unei distribuții, *semiinvarianții lui* THIELE: s_i [2].

Să efectuăm o schimbare de origine $y = x - a$. Vom avea:

$$\varphi_1(t) = \sum f_i e^{y_i t} = e^{-at} \sum f_i e^{x_i t} = e^{x-at} \varphi(t)$$

sau:

$$\varphi_1(t) = \psi(t) - at .$$

De aici deducem că toți semiinvarianții lui Thiele, cu singura excepție a lui s_1, rămân invariabili față de o deplasare a originii. Acesta a și fost motivul pentru care THIELE le-a dat numele de semiinvarianți.

---

[1] *Acest procedeu nu mai poate fi utilizat pentru demonstrarea aceleiași proprietăți în cazul distribuțiilor discontinue. Atunci însă, din cauza formei ei, funcția caracteristică satisface unei ecuații diferențiale liniare cu coeficienți constanți de ordinul s. Cunoașterea primelor s momente ne va permite, prin urmare, determinarea funcției caracteristice, care la rândul ei va conduce - după cum este ușor să ne asigurăm - printr-un sistem de ecuații algebrice, la determinarea completă a distribuției.*

[2] *Aceștia nu trebuie confundați cu semiinvarianții algebrici.*

Scanned with OKEN Scanner

Semiinvarianţii se pot obţine în funcţie de momente, prin identificarea coeficienţilor puterilor lui t din identitatea:

$$1+m_1\frac{t}{1!}+m_2\frac{t^2}{2!}+\ldots+m_p\frac{t^p}{p!}+\ldots=e^{s_1\frac{t}{1!}+s_2\frac{t^2}{2!}+\ldots+s_p\frac{t^p}{p!}+\ldots}$$

Obţinem în primul rând $s_1 = m_1$, ceea ce înseamnă că semi-invariantul de ordinul 1 este egal cu momentul de acelaşi ordin. Or, alegând ca origine media aritmetică a distribuţiei, noi putem face ca $m_1 = 0$; cum semiinvarianţii nu sunt schimbaţi prin această operaţie, identitatea precedentă devine:

$$1+\mu_2\frac{t^2}{2!}+\ldots+\mu_p\frac{t^p}{p!}+\ldots=e^{s_2\frac{t^2}{2!}+s_3\frac{t^3}{3!}+\ldots+s_p\frac{t^p}{p!}+\ldots}$$

formă sub care vom utiliza adesea funcţia caracteristică. Aplicând aici logaritmii şi identificând coeficienţii puterilor lui t, obţinem:

$$s_2=\mu_2 \qquad s_3=\mu_3 \qquad s_4=\mu_4-3\mu_2^2 \qquad s_5=\mu_5-10\mu_2\mu_3\ldots$$

relaţii care ne permit calculul imediat al semiinvarianţilor atunci când cunoaştem momentele.

Legătura dintre semiinvarianţi şi momente, de o parte, şi funcţia caracteristică, de alta, ne autorizează, în baza proprietăţii fundamentale a acesteia din urmă, să afirmăm că *şirul semiinvarianţilor sau al momentelor determină complet o distribuţie heterogradă*. Acesta este, de fapt, fondul teoretic al importanţei acestor caracteristici statistice.

210. *Compunerea variabilelor întâmplătoare.* Să considerăm două variabile întâmplătoare $X\begin{pmatrix}x_i\\f_i\end{pmatrix}$ şi $Y\begin{pmatrix}Y_j\\\varphi_j\end{pmatrix}$. Prin compunerea acestor variabile vom înţelege formarea unei a treia, $Z\begin{pmatrix}z_k\\\psi_k\end{pmatrix}$, în felul următor: fie (i_1, j_1), (i_2, j_2), $\ldots(i_r, j_r)$ cuplurile de valori pentru care $x_i + y_i = \text{const.} = z_k$. Dacă valorile celor două variabile sunt obţinute *independent* una de alta, atunci $\psi_k = \mathbf{S} f_i\varphi_j$, suma fiind extinsă asupra tuturor cuplurilor (i_a, j_a) menţionate[1].

[1] *Această definiţie este pur formală şi, până la un punct, arbitrară. Totuşi, pentru a o putea face utilă în practică, ea a fost construită pe baza celor două teoreme fundamentale ale calculului probabilităţilor (teoremele probabilităţilor totale şi compuse). Se vede imediat că Z îndeplineşte condiţia esenţială a variabilelor întâmplătoare:* $\sum\psi_k = 1$.

Fie acum $\psi(t)$ funcţia caracteristică a variabilei Z; vom avea:

$$\varphi(t) = \sum_{ki,j} \mathbf{S} f_i \varphi_j e^{(x_i+y_j)t} = \sum_{ki,j} \mathbf{S} f_i \varphi_j e^{x_i t} = \sum_{ki,j} \mathbf{S} f_i \varphi_j e^{y_j t} = f(t) \cdot \varphi(t)$$

unde f şi φ sunt respectiv funcţiile caracteristice ale variabilelor X şi Y. Acest rezultat se poate extinde la un număr oarecare de variabile şi constituie una din cele mai utile proprietăţi ale funcţiei caracteristice. Mai deducem că dacă s_i' şi s_i'' sunt semi-invarianţii distribuţiilor lui X şi Y, cei ai lui Z vor fi daţi de relaţia:

$$s_i = s_i' + s_i''$$

adică, *în compunerea variabilelor întâmplătoare, semiinvarianţii se adună.*

În particular, dacă dintr-o colectivitate teoretică $X\begin{pmatrix} x_i \\ f_i \end{pmatrix}$ a cărei funcţie caracteristică este $\varphi(t)$, facem n extracţii independente una de alta: $X_1, X_2, \ldots, X_n$, variabila $X_1 + X_2 + \ldots + X_n$ va avea ca funcţie caracteristică $\overline{\varphi(t)^n}$, iar funcţia caracteristică a mediei $\dfrac{X_1 + X_2 + \ldots + X_n}{n}$ va fi $\overline{\varphi\left(\dfrac{t}{n}\right)^n}$. Prin urmare, dacă s_i sunt semi-invarianţii unei distribuţii, cei ai distribuţiei mediei statistice a unui eşantion compus din n elemente vor fi:

$$\mathbf{s}_i = \frac{s_i}{n^{i-1}}$$

Această relaţie ne permite să determinăm distribuţia mediei statistice când cunoaştem distribuţia teoretică. Din relaţia stabilită deducem imediat $\mathbf{s}_1 = s_1$, rezultat care ne arată că media aritmetică a mediilor statistice este egală cu media aritmetică a colectivităţii teoretice (§ 102).

211. *Distribuţia binomială.* Din calculul probabilităţilor cunoaştem următorul rezultat:

Dacă dintr-o urnă conţinând bile albe şi negre în proporţia $p : q$ facem n extracţii, având grijă ca după fiecare extracţie să punem la loc bila extrasă, frecvenţele teoretice ale cazurilor în care au fost extrase respectiv 0, 1, 2 ..., n bile albe sunt:

$$q^n,\ C_n^1 q^{n-1} p,\ C_n^2 q^{n-2} p^2, \ldots\ldots, C_n^i q^{n-i} p^i, \ldots\ldots, p^n.$$

Funcţia caracteristică a acestei distribuţii se poate obţine fie direct, fie prin aplicarea rezultatelor referitoare la compunerea variabilelor întâmplătoare. Fiecare rezultat al unei extracţii este o variabilă întâmplătoare $X\begin{pmatrix} 0,1 \\ q,p \end{pmatrix}$, adică

poate fi 0 sau 1 după cum am scos o bilă albă sau nu. Funcţia caracteristică a lui X va fi evident $q + pe^t$, iar a rezultatului a n extracţii:

$$\varphi(t) = (q + pe^t)^n. \tag{164}$$

De aici deducem imediat $m_1 = p_n$; adoptând această valoare ca origine, funcţia (164) devine:

$$\varphi(t) = (q + pe^t)^n \cdot e^{-pnt} = (qe^{-pt} + pe^{qt})^n.$$

Se verifică imediat că:

$$\frac{\partial \varphi}{\partial q} = n(qe^{-pt} + pe^{qt})^{n-1}(e^{-pt} - e^{qt} + qte^{-pt} + pte^{qt}) =$$

$$= nt\varphi - \frac{1}{pq} \cdot \frac{\partial \varphi}{\partial t}$$

de unde, identificând coeficienţii puterilor lui t:

$$\mu_{i+1} = pq\left(ni\mu_{i-1} - \frac{\partial \mu_1}{\partial q}\right)$$

şi cum $\mu_0 = 1$, $\mu_1 = 0$, deducem:

$$\mu_2 = npq$$

$$\mu_3 = npq(q - p)$$

$$\mu_4 = npq(3npq + 1 - 6pq)$$

iar de aici (60):

$$\sqrt{\beta_1} = \frac{q - p}{\sqrt{npq}} \qquad \beta_2 = 3 + \frac{1 - 6pq}{npq}.$$

212. *Cazul eşantioanelor numeroase.* Rezultatele precedente sunt valabile oricare ar fi valoarea lui n. Este interesant de văzut ce formă va lua distribuţia binomială în cazul când n tinde către infinit. În acest scop, să considerăm poligonul respectiv care este determinat pentru valorile întregi ale lui r prin formula:

$$y_r = C_n^r q^{n-r} p^r.$$

Vom avea:

Scanned with OKEN Scanner

$$\Delta y_r = y_{r+1} - y_r = \frac{n!}{(n-r-1)!r!} q^{n-r-1} p^r \left(\frac{p}{r+1} - \frac{q}{n-r} \right)$$

și, notând prin:

$$y_m = \frac{y_{r+1} + y_r}{2} = \frac{n!}{2(n-r-1)!r!} q^{n-r-1} p^r \left(\frac{p}{r+1} - \frac{q}{n-r} \right)$$

ordonata medie a intervalului considerat, obținem:

$$\frac{\Delta y}{y_m} = \frac{2(np - q - r)}{np + q + r(q-p)}.$$

Vom lua ca origine abscisa medie $r = np$ și vom adopta pentru abscise o nouă unitate de măsură $\sigma = \sqrt{npq}$. Acest lucru implică - pentru ca proporționalitatea dintre suprafețe și frecvențe să fie conservată - înlocuirea lui y prin $\frac{y}{\sigma}$. Cu aceste schimbări, relația precedentă devine:

$$\frac{\Delta y}{y_m} = -\frac{\rho + \frac{q}{\sqrt{npq}}}{1 + \frac{1}{2np} + \frac{\sqrt{\beta_1}}{2}\rho} \cdot \frac{1}{\sqrt{npq}}$$

și, făcând o nouă schimbare de origine $\rho = x - \frac{q}{\sqrt{npq}}$:

$$\frac{\Delta y}{y_m} = -\frac{x}{1 + \frac{1}{n} + \frac{\sqrt{\beta_1}}{2}x} \cdot \frac{1}{\sqrt{npq}}.$$

Cum $\frac{1}{\sqrt{npq}}$ este un element infinitezimal în ipoteza că n tinde către infinit și cum panta poligonului care era dată de $\frac{\Delta y}{1}$ a devenit $\Delta y : \frac{1}{\sigma}$, vom avea la limită:

$$\frac{dy}{y} = -\frac{x}{1 + \frac{\sqrt{\beta_1}}{2}x} dx \qquad (165)$$

ecuația diferențială a limitei distribuției binomiale. Două cazuri se pot prezenta:

Scanned with OKEN Scanner

1. *p și q nu diferă prea mult între ei, astfel că* $\sqrt{\beta_1}$ *poate fi neglijat pentru n mare*. Ecuația (165) conduce la:

$$y = y_0 e^{-\frac{x^2}{2}}$$

unde y_0 este determinat prin condiția:

$$\int_{-\infty}^{+\infty} y dx = 1 \qquad y_0 = \frac{1}{\sqrt{2\pi}}.$$

Prin urmare, limita distribuției binomiale pentru eșantioanele numeroase și în condițiile menționate este chiar distribuția normală (§ 95). Revenind la vechea unitate de măsură, obținem:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}.$$

2. $\sqrt{\beta_1}$ *nu mai poate fi neglijat*. Ecuația (165) este de forma

$$\frac{dy}{y} = -\frac{\gamma x}{a + x} dx$$

care conduce la:

$$y = y_0 e^{-\gamma a}\left(1 + \frac{x}{a}\right)^{\gamma a}. \qquad (166)$$

Variabila variază de la -a la +∞, iar distribuția este ușor asimetrică. Aceasta este cunoscută sub numele de *distribuția de tipul III Pearson*. Constanta de integrare se determină prin condiția:

$$y_0 \int_a^{\infty} e^{-\gamma x}\left(1 + \frac{x}{a}\right)^{\gamma a} dx = 1$$

de unde:

$$y_0 = \left(\frac{\gamma a}{e}\right)^{\lambda a} \cdot \frac{\gamma}{\Gamma(\gamma a + 1)}.$$

Forma (166) este destul de incomodă pentru aplicații. Din această cauză s-a recurs la o aproximație a membrului al doilea:

Scanned with OKEN Scanner

$$y = y_0\left(1 - \frac{\gamma x}{1!} + \frac{\gamma^2 x^2}{2!} - \frac{\gamma^3 x^3}{3!} + \ldots\right)\left[1 + \frac{\gamma a}{1!}\cdot\frac{x}{a} + \frac{\gamma a(\gamma a - 1)}{2!}\cdot\frac{x^2}{a^2} + \ldots\right] =$$

$$= y_0\left[1 - \frac{\gamma}{2a}x^2 + \frac{\gamma}{3a^2}x^3 + \left(\frac{\gamma^2}{8a^4} - \frac{\gamma}{a^3}\right)x^4 \ldots\right] =$$

$$= y_0 e^{-\frac{\gamma x^2}{2a}}\left(1 + \frac{a\gamma}{3}\cdot\frac{x^3}{a^3} - \frac{a\gamma}{4}\cdot\frac{x^4}{a^4} + \ldots\right).$$

Cum în cazul nostru $a = \gamma = \frac{2}{\sqrt{\beta_1}}$, neglijând termenii de un ordin infinitezimal mai mare decât $\sqrt{\beta_1}$, avem:

$$y = y_0 e^{-\frac{x^2}{2}}\left(1 + \frac{\sqrt{\beta_1}}{6}x^3\right) \qquad y_0 = \frac{1}{\sqrt{2\pi}}. \tag{167}$$

Relaţia precedentă pune în evidenţă termenul corectiv care trebuie introdus pe lângă forma normală pentru a o face valabilă şi în cazul unei asimetrii pronunţate. Procedeul descris poate să fie împins la nevoie până la obţinerea unor termeni de grad mai mare în *x*.

Dacă *p* este de ordinul $\frac{1}{n}$, atunci *np* rămâne finit, iar demonstraţia de mai sus nu se mai poate face. Limita distribuţiei binomiale este în acest caz *legea lui Poisson* (§ 97).

Fie *np* = *m*; avem:

$$\frac{y_{r+1}}{y_r} = \frac{n-r}{r+1}\cdot\frac{p}{q} = \frac{1 - \frac{r}{n}}{q}\cdot\frac{m}{r+1}$$

la limită:

$$y_{r+1} = \frac{m}{r+1}\cdot y_r$$

de unde:

$$y_r = y_0\frac{m^r}{r!}$$

și cum, pe de altă parte, în cazul frecvențelor relative $\sum_0^\infty y_r = 1$, deducem:

$$y_0 = e^{-m}$$

deci:

$$y_r = e^{-m} \cdot \frac{m^r}{r!}.$$

Funcția caracteristică a acestei distribuții este:

$$\varphi(t) = \sum_0^\infty e^{rt} \cdot e^{-m} \cdot \frac{m^r}{r!} = e^{m(e^t - 1)}.$$

Aceasta ne pune în evidență caracteristica cea mai importantă a legii evenimentelor rare: *toți semi-invarianții unei astfel de distribuții sunt egali între ei.*

213. *Distribuțiile de tip Pearson.* Cazul mai complicat - dar în același timp mai reprezentativ pentru eșantioanele obținute în practică - când bila extrasă nu mai este repusă în urnă, conduce la distribuția:

$$\begin{matrix} 0, & 1 & , \ldots & \nu & , \ldots, & n \\ \frac{C_B^n}{C_N^n}, & \frac{C_A^1 C_B^{n-1}}{C_N^n}, & \ldots & \frac{C_A^\nu C_B^{n-\nu}}{C_N^n}, \ldots, & & \frac{C_A^n}{C_N^n} \end{matrix}$$

unde am însemnat prin N numărul bilelor din urnă, printre care A sunt albe, iar B negre.

Vom avea, cu notațiile din paragraful precedent:

$$\frac{\Delta y}{2y_m} = \frac{n(A+1) - (B+1) - (N+2)r}{n(A-1) + (B+1) + (B-A+2)r + 2r^2}$$

Pornind de la această formă; prof. K. PEARSON a definit clasa de distribuții care-i poartă numele și care sunt date de integrala ecuației diferențiale:

$$\frac{dy}{y} = \frac{(x-a)dx}{b_0 + b_1 x + b_2 x^2} \qquad (168)$$

cunoscută sub numele de *ecuația celor patru constante*. Ea conduce, după natura rădăcinilor trinomului $b_0 + b_1 x + b_2 x^2$, la șapte tipuri mai importante de distribuții, a căror caracteristică principală este faptul că sunt unimodale.

Scanned with OKEN Scanner

(Abscisa modului este dată de ecuația $\frac{dy}{dx} = 0$). Aplicațiile practice ale tipurilor Pearson au arătat că acestea pot reprezenta într-un mod satisfăcător aproape toate distribuțiile ce se pot întâlni. Pentru acest motiv, orice generalizare a formei (168), în afară de faptul că ar introduce complicații analitice destul de serioase, pare de prisos.

Din relația (168) se deduce:

$$\left(b_0 x^p + b_1 x^{p+1} + b_2 x^{p+2}\right) dy = x^p (x - a) y\, dx$$

Integrând prin părți și presupunând că la limitele intervalului de variație $x^n y$ este nul, vom avea:

$$-b_0 p \int x^{p-1} y\, dx - b_1 (p+1) \int x^p y\, dx - b_2 (p+2) \int x^{p+1} y\, dx =$$
$$\int x^{p+1} y\, dy - a \int x^p y\, dy$$

sau:

$$b_0 p m_{p-1} + \left[b_1 (p+1) - a\right] m_p + \left[b_2 (p+2) + 1\right] m_{p+1} = 0. \qquad (169)$$

Presupunând că originea este chiar media aritmetică, această relație conduce la sistemul:

$$\begin{aligned} & b_1 - a = 0 \\ & b_0 + 3 b_2 \mu_2 = -\mu_2 \\ & 3 b_1 \mu_2 + 4 b_2 \mu_3 - a \mu_2 = -\mu_3 \\ & 3 b_0 \mu_2 + 4 b_1 \mu_3 + 5 b_2 \mu_4 - a \mu_3 = -\mu_4 \end{aligned}$$

de unde se deduce:

$$a = b_1 = -\frac{\sqrt{\mu_2}\sqrt{\beta_1}(\beta_2 + 3))}{D} \qquad b_0 = \frac{\mu_2 (3\beta_1 - 4\beta_2)}{D}$$

$$b_2 = \frac{6 + 3\beta_1 - 2\beta_2}{D} \qquad D = 2(5\beta_2 - 6\beta_1 - 9).$$

Prin urmare, pentru tipurile Pearson, primele patru momente determină complet distribuția. Celelalte momente se obțin prin relația de recurență (169). Ca o aplicație imediată a acestei proprietăți, vom considera tipul III Pearson, pentru care:

Scanned with OKEN Scanner

$$a = b_2 = 0 \qquad b_0 = -\frac{a}{\gamma} \qquad b_1 = -\frac{1}{\gamma}.$$

Momentul de ordinul 1 se obține făcând în (169) $p = 0$:

$$m_1 = \frac{1}{\gamma}$$

iar, în urmă, prin recurență:

$$m_2 = \frac{a\gamma + 2}{\gamma^2} \qquad m_3 = \frac{5a\gamma + 6}{\gamma^3} \qquad m_4 = \frac{3a^2\gamma^2 + 26a\gamma + 24}{\gamma^4}$$

de unde:

$$\mu_2 = \frac{a\gamma + 1}{\gamma^2} \qquad \mu_3 = \frac{2(a\gamma + 1)}{\gamma^3} \qquad \mu_4 = \frac{3(a\gamma + 1)(a\gamma + 3)}{\gamma^4}$$

$$\beta_1 = \frac{4}{a\gamma + 1} \qquad \beta_2 = 3 + \frac{6}{a\gamma + 1}.$$

Toate aceste valori ar fi putut fi obținute și prin integrare directă, însă calculele ar fi fost în acest caz mult mai laborioase[1].

Din ultimele egalități deducem:

$$2\beta_2 - 3\beta_1 - 6 = 0$$

relație care caracterizează acest tip de distribuție.

În general, forma distribuțiilor depinde de natura rădăcinilor trinomului $b_0 + b_1x + b_2x^2$ care la rândul lor depind de valorile expresiilor:

$$k_1 = -b_2D = 2\beta_2 - 3\beta_1 - 6$$

$$k_2 = \frac{b_1^2}{4b_0b_2} = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}.$$

Prof. K. Pearson a făcut o amplă analiză a tuturor cazurilor ce se pot prezenta. Principalele rezultate obținute sunt concretizate în următorul tabel:

$k_2 = -\infty$	$k_2 < 0$	$k_2 = 0$	$0 < k_2 < 1$	$k_2 = 1$	$k_2 > 1$	$k_2 = +\infty$
Tip III	Tip I	Tip VII ($\beta_2 = 3$) Tip II ($\beta_2 < 3$)	Tip IV	Tip V	Tip VI	Tip III

[1] *Pe aceeași cale se obțin relațiile (50).*

TIP I	$y = y_0\left(1+\frac{x}{a_1}\right)^{va_1}\left(1-\frac{x}{a_2}\right)^{va_2}$	*Rădăcini reale de semn contrar*
TIP II	$y = y_0\left(1-\frac{x^2}{a^2}\right)^m$	*Rădăcini reale egale*
TIP III	$y = y_0 e^{-\gamma x}\left(1+\frac{x}{a}\right)^{\gamma a}$	*O rădăcină infinită*
TIP IV	$y = y_0 e^{-v\,\mathrm{arctg}\,x}\left(1+\frac{x^2}{a^2}\right)^{-m}$	*Rădăcini imaginare*
TIP V	$y = y_0 x^{-p} e^{-\frac{\gamma}{x}}$	*Rădăcini nule*
TIP VI	$y = y_0 (x-a)^{q_2} x^{-q_1}$	*Rădăcini reale de acelaşi semn*
TIP VII	$y = y_0 e^{-\frac{x^2}{2\sigma^2}}$	*Rădăcini infinite*

Pe lângă acestea, mai există şi alte tipuri, numite *tipuri de tranziţie*, cum este, de pildă, distribuţia:

$$y = y_0\left(1+\frac{x^2}{a^2}\right)^{-m}$$

pentru care $k_2 = 0$, $\beta_2 > 3$ şi care se găseşte la limită între tipul VII şi tipul IV[1].

214. *Distribuţia normală*. Funcţia caracteristică a acestei distribuţii este:

$$\varphi(t) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{xt} e^{-\frac{x^2}{2\sigma^2}} dx = \frac{e^{\frac{t^2\sigma^2}{2}}}{\sigma\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{1}{2\sigma^2}(x-t\sigma)^2} dx = e^{\frac{t^2\sigma^2}{2}}$$

de unde deducem imediat

$$\mu_{2s+1} = 0 \qquad \mu_{2s} = \frac{(2s)!}{s!}\left(\frac{\sigma^2}{2}\right)^s. \tag{170}$$

[1] *Pentru o descriere completă a tuturor tipurilor, a se vedea W. Palin Elderton,* ***Frequency Curves and Correlation****, London, Layton, 1927.*

Cea mai importantă caracteristică a distribuției normale este faptul că *toți semiinvarianții de ordin superior lui 2 sunt nuli.* Însăși importanța teoretică a semiinvarianților decurge din această proprietate.

Să facem o aplicație imediată a acestui rezultat. Media statistică a unui eșantion extras dintr-o colectivitate normală va avea ca funcție caracteristică (§ 210):

$$\varphi_{M'}(t) = e^{\frac{t^2\sigma^2}{2n}} \tag{171}$$

căreia îi corespunde distribuția normală:

$$\frac{\sqrt{n}}{\sigma\sqrt{2\pi}} e^{-\frac{nM'^2}{2\sigma^2}}.$$

Acest rezultat se menține și pentru o distribuție oarecare, însă numai în cazul eșantioanelor numeroase; într-adevăr, fie:

$$\varphi(t) = e^{s_2\frac{t^2}{2!}+s_3\frac{t^3}{3!}+\dots}$$

funcția caracteristică a distribuției teoretice. Cea a mediei statistice va fi

$$\varphi_{M'}(t) = e^{\frac{s_2}{n}\cdot\frac{t^2}{2!}+\frac{s_3}{n^2}\cdot\frac{t^3}{3!}+\dots}$$

și, prin urmare, dacă n este suficient de mare, pentru a putea neglija termenii de un ordin infinitezimal mai mare decât $1/n$, vom obține relația (171).

215. *Modul.* O evaluare a modului se obține prin interpolare parabolică. Fie $x_i - d$, x_i, $x_i + d$ mijloacele a trei intervale, x_i fiind cel al intervalului modal, iar n_{i-1}, n_i, n_{i+1} frecvențele corespunzătoare. Luând pe x_i ca origine, parabola de ordinul 2 care trece prin punctele corespunzătoare, vârfuri ale poligonului de frecvențe, va avea ca ecuație:

$$\begin{vmatrix} y^2 & x^2 & x & 1 \\ n_{i-1} & d^2 & -d & 1 \\ n_i & 0 & 0 & 1 \\ n_{i+1} & d^2 & d & 1 \end{vmatrix} = 0$$

Maximumul acesteia va corespunde abscisei:

$$-\frac{d(n_{i+1}-n_{i-1})}{2(n_{i+1}-2n_i+n_{i-1})}$$

şi deci modul corect va fi (11):

$$Mo = x_i - \frac{n_{i+1}-n_{i-1}}{n_{i+1}-2n_i+n_{i-1}} \cdot \frac{d}{2}.$$

În cazul distribuţiilor de tip Pearson putem determina destul de uşor chiar modul adevărat. Am văzut mai sus că în cazul când originea este egală cu media aritmetică:

$$a = -\frac{\sqrt{\mu_2}\sqrt{\beta_1}(\beta_2+3)}{2(5\beta_2-6\beta_1-9)}.$$

Cum, pe de altă parte, *a* este abscisa modului distribuţiei, vom avea:

$$Mo = M - \frac{\sigma\sqrt{\beta_1}(\beta_2+3)}{2(5\beta_2-6\beta_1-9)},$$

de unde deducem relaţia (36).

Referindu-ne acum la distribuţia (167), modul va avea ca valoare $Mo = 0$. Media aritmetică a aceleiaşi distribuţii va fi în unităţi σ:

$$M = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} xe^{-\frac{x^2}{2}}\left(1+\frac{\sqrt{\beta_1}}{6}x^3\right)dx = \frac{\sqrt{\beta_1}}{2},$$

iar valoarea medianei va rezulta din rezolvarea ecuaţiei:

$$\int_{-\infty}^{Me} ydx = \int_{Me}^{+\infty} ydx$$

care se mai poate scrie:

$$\int_{-\infty}^{Me} e^{-\frac{x^2}{2}}dx - \frac{\sqrt{\beta_1}}{6}(Me^2+2)e^{-\frac{Me^2}{2}} = \int_{Me}^{+\infty} e^{-\frac{x^2}{2}}dx + \frac{\sqrt{\beta_1}}{6}(Me^2+2)e^{-\frac{Me^2}{2}}$$

De aici, cum:

$$\int_{0}^{Me} e^{-\frac{x^2}{2}}dx \approx Me \times e^{-\frac{Me^2}{2}}$$

obținem:

$$\sqrt{\beta_1}Me^2 - 6Me + 2\sqrt{\beta_1} = 0$$

și, prin urmare:

$$Me = \frac{3 - \sqrt{9 - 2\beta_1}}{\sqrt{\beta_1}} = \frac{1 - \left(1 - \frac{\beta_1}{9} + \dots\right)}{\frac{\sqrt{\beta_1}}{3}} = \frac{\sqrt{\beta_1}}{3} + \dots$$

Comparând între ele cele trei valori tipice obținute mai sus, obținem:

$$M - Mo = 3\,(M - Me)$$

relație care este utilizată pentru calculul aproximativ al modului distribuțiilor ușor asimetrice (12).

216. *Distribuții de mai multe variabile.* Acestea se definesc la fel ca și distribuțiile de o singură variabilă. Pentru cazul distribuțiilor discontinue este suficient să considerăm frecvența $n_{ij...l}$ a ansamblului de valori simultane $X = x_i$, $Y = y_j, \dots, U = u_l$, iar în cazul general distribuția va fi determinată prin *funcția frecvențelor cumulate* $N(x, y, \dots, u)$ care reprezintă frecvența cazurilor $X < x$, $Y < y, \dots, U < u$. Funcția $N(x, y, \dots, u)$ se bucură de aceleași proprietăți ca și $N(x)$ (§ 204). În particular, dacă derivata $\frac{\partial^\alpha N}{\partial x \partial y \dots\dots\dots \partial u} = f(x,y,\dots,u)$ există, frecvența cazurilor:

$$x \le X \le x + dx, \quad y \le Y \le y + dy, \dots, u \le U \le u + du$$

va fi egală cu $f dx dy \dots du$ și va reprezenta un element de volum al *suprafeței de corelație* $t = f(x, y, \dots, u)$.

Să ne fixăm ideile asupra unei distribuții de două variabile. Oricare din aceste variabile are o *distribuție liberă* care oglindește numai frecvențele diferitelor valori ale ei, independent de valorile pe care le ia cealaltă. Astfel, distribuția liberă a lui X va fi exprimată, după caz, prin frecvențele:

$$g(x) = \int f(x,y)dy \qquad n_{i|} = \sum_j n_{ij}$$

iar a lui Y:

$$h(y) = \int f(x,y)dx \qquad n_{|j} = \sum_i n_{ij}$$

Dacă X rămâne în intervalul $(x_i, x_i + dx_i)$, Y poate lua diferite valori care conduc la o *distribuţie legată* a acestei din urmă variabile. Distribuţia legată are ca funcţie de frecvenţă $f(x_i, y)$ sau, în frecvenţe relative:

$$\frac{f(x_i, y)}{g(x_i)}.$$

În raport cu o distribuţie de două variabile putem defini momentul produs de ordinul $p + q$ prin relaţia

$$m_{pq} = \iint x^p y^q f(x,y)dx\,dy.$$

Acesta se confundă cu unul din momentele libere dacă p sau q sunt nuli.

În afară de momentele-produs, putem calcula şirul complet al momentelor pentru fiecare distribuţie legată în parte:

$$m_{0q}^{(i)} = \frac{1}{g(x_i)} \int y^q f(x_i, y)dy$$

de unde:

$$m_{pq} = \int x_i^p m_{0q}^{(i)} g(x_i)dx_i \tag{172}$$

sau pentru variabile discontinue:

$$m_{pq} = \frac{\sum n_{i|} x_i^p m_{0q}^{(i)}}{N}. \tag{172 bis}$$

Funcţia caracteristică a unei distribuţii de două variabile se defineşte prin relaţia

$$\varphi(t,u) = \iint e^{tx+uy} f(x,y)dx\,dy. \tag{173}$$

De aici obţinem prin inversiune legea distribuţiei:

$$f(x,y) = \int e^{-i(Tx+Uy)} \Phi(T,U)dT\,dU \tag{174}$$

unde $\Phi(T,U) = \varphi(iT,iU)$.

Dacă funcţia caracteristică este dezvoltabilă în serie, atunci (173) generează momentele:

$$\varphi(t,u) = 1 + \sum\sum m_{pq} \frac{t^p}{p!} \cdot \frac{u^q}{q!} = 1 + \sum_1^\infty \frac{1}{s!}\left(m_{10}t + m_{01}u\right)^{(s)}\ ^1.$$

[1] *Puterea este simbolică.*

Funcțiile caracteristice ale distribuțiilor libere sunt respectiv $\varphi(t,0)$ și $\varphi(0,u)$.

Semiinvarianții distribuțiilor de două variabile se definesc prin procedeul cunoscut:

$$\psi(t,u) = \log\varphi(t,u) = \sum\sum s_{pq}\frac{t^p}{p!}\cdot\frac{t^q}{q!} = \sum_1^\infty \frac{1}{s!}\left(s_{10}t + s_{01}u\right)^{(s)}.$$

În afară de semiinvarianții de ordinul 1:

$$s_{10} = m_{10} \qquad s_{01} = m_{01}$$

toți ceilalți sunt invariabili față de orice schimbare a originii.

Adoptând pe m_{10} și m_{01} ca origini, obținem:

$$\log\left\{1 + \sum_2^\infty \frac{1}{s!}\left(\mu_{10}t + \mu_{01}u\right)^{(s)}\right\} = \sum_2^\infty \frac{1}{s!}\left(s_{10}t + s_{01}u\right)^{(s)}$$

de unde, ca mai sus (§ 209):

$$\left(s_{10}t + s_{01}u\right)^{(2)} = \left(\mu_{10}t + \mu_{01}u\right)^{(2)}$$
$$\left(s_{10}t + s_{01}u\right)^{(3)} = \left(\mu_{10}t + \mu_{01}u\right)^{(3)}$$
$$\left(s_{10}t + s_{01}u\right)^{(4)} = \left(\mu_{10}t + \mu_{01}u\right)^{(4)} - 3\left[\left(\mu_{10}t + \mu_{01}u\right)^{(2)}\right]^2$$

de unde:

$$s_{20} = \mu_{20} \qquad s_{11} = \mu_{11}$$
$$s_{30} = \mu_{30} \qquad s_{21} = \mu_{21}$$
$$s_{40} = \mu_{40} - 3\mu_{20}^2 \qquad s_{31} = \mu_{31} - 3\mu_{20}\mu_{11} \qquad s_{22} = \mu_{22} - 2\mu_{11}^2 - \mu_{20}\mu_{02}$$

La fel, din identitatea:

$$1 + \sum_2^\infty \frac{1}{s!}\left(\mu_{10}t + \mu_{01}u\right)^{(s)} = \left[1 + \sum_1^\infty \frac{1}{s!}\left(m_{10}t + m_{01}u\right)^{(s)}\right] \times e^{-(m_{10}t + m_{01}u)}$$

obținem:

Scanned with OKEN Scanner

$$(\mu_{10}t + \mu_{01}u)^{(2)} = (m_{10}t + m_{01}u)^{(2)} - (m_{10}t + m_{01}u)^{(2)}$$

$$(\mu_{10}t + \mu_{01}u)^{(3)} = (m_{10}t + m_{01}u)^{(3)} - 3(m_{10}t + m_{01}u)^{(2)}(m_{10}t + m_{01}u) + 2(m_{10}t + m_{01}u)^{(3)}$$

. . . . . .

sau:

$$\mu_{20} = m_{20} - m_{10}^2 \qquad \mu_{11} = m_{11} - m_{10}m_{01}$$

$$\mu_{30} = m_{30} - 3m_{20}m_{10} + 2m_{10}^3 \qquad \mu_{21} = m_{21} - m_{20}m_{01} - 2m_{11}m_{10} + 2m_{10}^2 m_{01}$$

. . . . . .

217. *Legătura stocastică.* Dacă valorile lui X nu influenţează întru nimic valorile pe care le ia Y, există independenţă completă între cele două variabile. În acest caz, distribuţiile legate vor fi identice între ele şi identice cu distribuţiile libere corespunzătoare, adică vom avea:

$$\frac{f(x_i, y)}{g(x_i)} = h(y).$$

Prin urmare, funcţia de frecvenţă se va descompune într-un produs de două funcţii de câte o singură variabilă:

$$f(x,y) = g(x)\, h(y). \tag{175}$$

Aceasta este condiţia necesară şi suficientă a independenţei variabilelor. Identitatea (175) conduce la relaţia:

$$\varphi(t,u) = \varphi(t,0)\, \varphi(0,u) \tag{176}$$

de unde deducem o altă formă a condiţiei necesare şi suficiente pentru independenţa variabilelor X şi Y:

$$s_{pq} = 0 \quad \text{când} \quad pq = 0. \tag{177}$$

În orice alt caz variabilele sunt legate stocastic între ele. Distribuţiile legate ale lui X vor varia în raport cu valorile lui Y şi viceversa. Orice distribuţie fiind determinată prin momentele respective, legătura stocastică se va traduce printr-o serie de funcţii $m_{01}^{(i)}(x)$ şi $m_{10}^{(j)}(y)$. S-ar putea totuşi ca unele din aceste funcţii să fie constante. Astfel, să considerăm cazul: $m_{01}^{(i)} = \text{const}$. Cu ajutorul relaţiei (172), deducem:

$$m_{p1} = m_{01}^{(i)} m_{p0}$$

Scanned with OKEN Scanner

de unde $m_{01}^{(i)} = m_{01}$ *şi* $\mu_{p1} = 0$.

Se obţine imediat:

$$s_{p1} = 0 \tag{178}$$

condiţie care nu acoperă toate posibilităţile cuprinse în (177). Variabilele îndeplinind condiţia (178), nu sunt, prin urmare, în general independente. Vom spune în acest caz că Y este necorelat cu X (§ 142). După cum vedem, variabilele necorelate nu sunt teoretic caracterizate numai prin relaţia $\mu_{11} = 0$.

Să presupunem acum că $\mu_{02}^{(i)}$ = const. O astfel de distribuţie poartă numele de *homoscedastică* în Y. *Dacă pe deasupra Y este necorelat cu X*, vom avea $m_{02}^{(i)}$ = const. şi deci (172):

$$m_{p2} = m_{02}^{(i)} m_{p0}$$

de unde:

$$m_{02}^{(i)} = m_{02} \qquad \mu_{02}^{(i)} = \mu_{02} \qquad \mu_{p2} = \mu_{02}\mu_{p0}.$$

De aici se deduce:

$$s_{p2} = 0 \tag{179}$$

relaţie care ne prilejuieşte aceleaşi constatări ca şi (178).

218. *Regresie*. Să considerăm curba reprezentată de funcţia $\tilde{y}_x = m_{01}^{(i)}$. Ea poartă, după cum ştim, numele de linie de regresie a lui Y în X. În cazul unei regresii *liniare*, vom avea:

$$m_{01}^{(i)} = a + b_{21}x.$$

Introducând în (172), obţinem:

$$m_{p1} = am_{p0} + b_{21}m_{p+1,0} \tag{180}$$

de unde:

$$m_{01} = a + b_{21}m_{10}$$
$$m_{11} = am_{10} + b_{21}m_{20}$$

Presupunând că originile au fost convenabil alese (§ 137), sistemul precedent ne dă:

Scanned with OKEN Scanner

$$a = 0 \qquad b_{21} = \frac{\mu_{11}}{\mu_{20}}$$

iar relația (180) devine:

$$\mu_{p1}\mu_{20} = \mu_{11}\mu_{p+1,0}. \tag{181}$$

Aceasta este condiția necesară ca regresia lui Y în X să fie liniară. Condiția de liniaritate a regresiei lui X în Y diferă de (181) și, prin urmare, dacă una din regresii este liniară, nu urmează neapărat că și cealaltă va fi la fel.

Dacă avem și homoscedasticitate în Y:

$$m_{02}^{(i)} = \frac{\mu_{11}^2}{\mu_{20}^2} x^2 + \mu_{02}^{(i)}$$

deci din (172):

$$\mu_{p2} = \mu_{02}^{(i)}\mu_{p0} + \frac{\mu_{11}^2}{\mu_{20}^2}\mu_{p+2,0}$$

de unde, pentru $p = 0$ (§ 150):

$$\mu_{02}^{(i)} = \mu_{02}(1 - r^2). \tag{182}$$

Dacă linia de regresie este o parabolă de ordinul 2:

$$m_{01}^{(i)} = a_0 + a_1 x + a_2 x^2$$

deducem, ca mai sus:

$$\mu_{p1} = a_0\mu_{p0} + a_1\mu_{p+1,0} + a_2\mu_{p+2,0}$$

de unde:

$$\begin{aligned} a_0 + \mu_{20}a_2 &= 0 \\ \mu_{20}a_1 + \mu_{30}a_2 &= \mu_{11} \\ \mu_{20}a_0 + \mu_{30}a_1 + \mu_{40}a_2 &= \mu_{21} \end{aligned}$$

sistem care conduce la valorile

$$a_0 = \frac{\mu_{20}\left(\mu_{11}\mu_{30} - \mu_{21}\mu_{20}\right)}{\mu_{20}\mu_{40} - \mu_{30}^2 - \mu_{20}^3} \qquad a_1 = \frac{\mu_{11}}{\mu_{20}} - a_2\frac{\mu_{30}}{\mu_{20}} \qquad a_2 = -\frac{a_0}{\mu_{20}}.$$

Pentru cazul când linia de regresie este exact sau aproximativ exprimată printr-un polinom oarecare, determinarea ei se face mai ușor prin metoda poligoanelor ortogonale descrisă de TCHEBICHEFF. Fie:

Scanned with OKEN Scanner

$$P_0(x), \quad P_1(x), \quad P_2(x), \quad \ldots P_k(x), \ldots\ldots$$

un şir de poligoane ortogonale în raport cu distribuţia liberă a lui X, indicele fiind egal cu gradul polinomului respectiv. Condiţia de ortogonalitate este:

$$\sum_i n_{i|} P_k(x_i) P_j(x_i) = 0 \qquad k \neq j. \tag{183}$$

Luând $P_0 = 1$ şi $P_1 = x + a_0^1$, vom avea, în baza acestei condiţii:

$$\sum n_{i|} x_i + N a_0^1 = 0$$

şi dacă $m_{10} = 0$, obţinem $P_1 = x$. La fel $P_2 = x^2 + a_1^2 x + a_0^2$. Condiţiile de ortogonalitate vor fi:

$$\mu_{20} + a_0^2 = 0 \qquad \mu_{30} + a_1^2 \mu_{20} = 0$$

deci:

$$P_2 = x^2 - \frac{\mu_{30}}{\mu_{20}} x - \mu_{20}.$$

În general:

$$P_k(x) = \frac{(-1)^k}{\Delta} \begin{vmatrix} 1 & x & x^2 & x^k \\ & & & \mu_{k_0} \\ & & & \mu_{k+1,0} \\ & & \Delta & \cdot \\ & & & \cdot \\ & & & \cdot \\ & & & \mu_{2k-1,0} \end{vmatrix}$$

unde:

$$\Delta = \begin{vmatrix} 1 & 0 & \mu_{20} \cdots & \mu_{k-1,0} \\ 0 & \mu_{20} & \mu_{30} \cdots & \mu_{k,0} \\ \mu_{20} & \mu_{30} & \mu_{40} \cdots & \mu_{k+1,0} \\ \mu_{k-1,0} & \mu_{k,0} & \mu_{k+1,0} \cdots & \mu_{2k-2,0} \end{vmatrix}.$$

Scanned with OKEN Scanner

Punând:

$$m_{01}^{(i)} = \alpha_0 + \alpha_1 P_1 + \alpha_2 P_2 + \ldots + \alpha_s P_s,$$

se deduce, în baza proprietății de ortogonalitate:

$$\alpha_k = \frac{\sum n_{i|} m_{01}^{(i)} P_k(x_i)}{\sum n_{i|} \overline{P_k(x_i)}^2}$$

relație care determină toți coeficienții α_k. Pentru $s = 2$, cazul tratat mai sus:

$$\alpha_0 = 0 \qquad \alpha_1 = \frac{\mu_{11}}{\mu_{02}} \qquad \alpha_2 = \frac{\mu_{20}\mu_{21} - \mu_{11}\mu_{30}}{\mu_{20}\mu_{40} - \mu_{30}^2 - \mu_{20}^3}$$

iar:

$$m_{01}^{(i)} = \frac{\mu_{11}}{\mu_{02}} x + \frac{\mu_{20}\mu_{21} - \mu_{11}\mu_{30}}{\mu_{20}\mu_{40} - \mu_{30}^2 - \mu_{20}^3} \left(x^2 - \frac{\mu_{30}}{\mu_{20}} x - \mu_{20} \right)$$

adică același rezultat, în fond, ca cel obținut pe prima cale. Metoda lui TCHEBICHEFF are însă un mare avantaj: *calculele făcute rămân valabile chiar dacă se introduce în expresia lui* $m_{01}^{(i)}$ *un nou termen de un grad mai înalt*, deoarece fiecare coeficient α_k nu depinde decât de polinomul respectiv P_k [1].

219. *Raportul de corelație*. Acesta este definit prin formula (113):

$$\eta_{yx}^2 = 1 - \frac{1}{\mu_{02}} \cdot \frac{\sum n_{i|} \mu_{02}^{(i)}}{N}.$$

Cum:

$$\sum_i n_{i|} \mu_{02}^{(i)} = \sum_i \sum_j n_{ij} \left(y_j - m_{01}^{(i)} \right)^2 = \sum_i \sum_j n_{ij} \left(y_j - m_{01} + m_{01} - m_{01}^{(i)} \right)^2 =$$

$$= \sum_i \sum_j n_{ij} \left(y_j - m_{01} \right)^2 - \sum n_{i|} \left(m_{01} - m_{01}^{(i)} \right)^2$$

relația precedentă devine (§ 151):

[1] *Aceeași metodă este utilizată* mutatis mutandis *și pentru interpolarea parabolică a tendinței de lungă durată în cazul seriilor în timp.*

Scanned with OKEN Scanner

$$\eta_{yx}^2 = \frac{1}{\mu_{02}} \cdot \frac{\sum n_{i|}\left(m_{01} - m_{01}^{(i)}\right)^2}{N}.$$

Să punem acum:

$$Y_i = \frac{\mu_{11}}{\mu_{20}} x_i = b_{21} x_i$$

și să considerăm $m_{01} = 0$; vom avea:

$$\sum n_{i|}\left(m_{01}^{(i)}\right)^2 = \sum n_{i|}\left(m_{01}^{(i)} - Y_i + Y_i\right)^2$$

$$= \sum n_{i|}\left(m_{01}^{(i)} - Y_i\right)^2 + N\frac{\mu_{11}^2}{\mu_{20}}$$

deci:

$$\eta_{yx}^2 - r^2 = \frac{1}{\mu_{02}} \cdot \frac{\sum n_{i|}\left(m_{01}^{(i)} - Y_i\right)^2}{N}.$$

Prin urmare, diferența $\eta_{yx}^2 - r^2$ este pozitivă și nu se anulează decât în cazul regresiei perfect lineare.

220. *Corelația normală*. Distribuția normală de o singură variabilă a fost introdusă ca limita distribuției binomiale. Vom considera acum o urnă în care există trei feluri de bile: albe, roșii și negre, în proporțiile p_1, p_2, p_3. Vom avea evident $p_1 + p_2 + p_3 = 1$. Urmărind numai culorile albă și roșie, o extracție va putea avea ca rezultat una din eventualitățile:

$$\begin{matrix} X & \begin{pmatrix}1\\0\end{pmatrix} & \begin{pmatrix}0\\1\end{pmatrix} & \begin{pmatrix}0\\0\end{pmatrix} \\ Y & & & \end{matrix}$$

respectiv cu frecvențele p_1, p_2, p_3. Prin urmare, funcția caracteristică a unei extracții va fi $p_1 e^{t_1} + p_2 e^{t_2} + p_3$, iar aceea a n extracții:

$$\varphi(t_1, t_2) = \left(p_1 e^{t_1} + p_2 e^{t_2} + p_3\right)^n.$$

Se deduce imediat $m_{10} = np_1$ și $m_{02} = np_2$. Adoptând aceste valori ca origini, vom avea:

Scanned with OKEN Scanner

$$\varphi(t_1,t_2) = e^{-n(p_1t_1+p_2t_2)}\left(p_1e^{t_1} + p_2e^{t_2} + p_3\right)^n$$

de unde, punând $1 - p_i = q_i$, obținem:

$$\psi(t_1t_2) = \log\varphi(t_1t_2) = n\left(p_1\frac{t_1^2}{2!} + p_2\frac{t_2^2}{2!} - \frac{(p_1t_1 + p_2t_2)^2}{2} + \ldots\right)$$

$$= n\left(p_1q_1\frac{t_1^2}{2!} - p_1p_2\frac{t_1}{1!}\frac{t_2}{1!} + p_2q_2\frac{t_2^2}{2!} + \ldots\right)$$

și, prin urmare:

$$\mu_{20} = np_1q_1 \qquad \mu_{11} = -np_1p_2 \qquad \mu_{02} = np_2q_2. \tag{184}$$

Să măsurăm abaterile față de valorile medii np_1, np_2 în unități σ_1 și σ_2 și să neglijăm termenii conținând pe n la numitor[1]; funcția caracteristică devine:

$$\Phi(t_1,t_2) = e^{\frac{1}{2}\left(t_1^2 + 2rt_1t_2 + t_2^2\right)} \tag{185}$$

unde:

$$r = \frac{\mu_{11}}{\sigma_1\sigma_2} = -\sqrt{\frac{p_1p_2}{q_1q_2}}$$

Acesteia îi corespunde distribuția:

$$f(X_1X_2) = \frac{1}{4\pi^2}\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-i\left(t_1X_1 + t_2X_2\right)}e^{-\frac{1}{2}\left(t_1^2 + 2rt_1t_2 + t_2^2\right)}dt_1dt_2$$

$$= \frac{1}{2\pi\sqrt{1-r^2}}e^{-\frac{1}{2(1-r^2)}\left(X_1^2 - 2rX_1X_2 + X_2^2\right)} \tag{186}$$

care poartă numele de *distribuție normală* de două variabile (§146).

Revenind la vechile unități de măsură, distribuția normală se înfățișează sub forma[2]:

---

[1] *Pentru aproximații mai bune, a se vedea G. Darmois,* ***Statistique mathématique****, Doin, Paris, 1928, p. 240.*

[2] *Nu trebuie uitat, cu prilejul acestor schimbări de unități de măsură, că frecvența cazurilor $x \leq X \leq x + dx$, $y \leq Y \leq y + dy$ este de fapt $f(x,y)dx\,dy$ și că, prin urmare, trebuie să ținem seama și de transformarea elementului $dx\,dy$.*

Scanned with OKEN Scanner

$$f(x_1x_2)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}e^{-\frac{1}{2(1-r^2)}\left(\frac{x_1^2}{\sigma_1^2}-2r\frac{x_1}{\sigma_1}\cdot\frac{x_2}{\sigma_2}+\frac{x_2^2}{\sigma_2^2}\right)}.$$

Relația (185) ne permite următoarele constatări:

1. Variabilele X, Y nu vor putea fi necorelate decât dacă r (respectiv μ_{11}) va fi nul (178). În acest din urmă caz, ele vor fi însă și complet independente (177). Iată cum se dovedește că pentru distribuția normală lipsa corelației este echivalentă cu independența absolută a variabilelor (§ 146).

2. Ambele regresii sunt liniare (181). Deci:

$$\eta_{yx}^2=\eta_{xy}^2=r^2.$$

3. Distribuția normală este homoscedastică în X și Y. Valabilitatea relației (182) se poate constata și direct:

$$g(x)=\frac{1}{\sigma_x\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma_x^2}}$$

$$\frac{f(x_i,y)}{g(x_i)}=\frac{1}{\sqrt{2\pi}\sqrt{\sigma_y(1-r^2)}}e^{-\frac{1}{2\sigma_y^2(1-r^2)}\left(y-r\frac{\sigma_y}{\sigma_x}x_i\right)^2}.$$

Cu alte cuvinte, distribuțiile legate sunt normale, iar fluctuația lor în jurul mediilor respective $r\frac{\sigma_y}{\sigma_x}\cdot x_i$ este egală cu $\sigma_y^2\left(1-r^2\right)$.

Pătratul mediu de contingență al unei distribuții normale va fi:

$$\varphi^2=\iint\frac{\left[\frac{1}{2\pi\sqrt{1-r^2}}e^{-\frac{1}{2}\alpha(x,y)}-\frac{1}{2\pi}e^{-\frac{1}{2}\beta(x,y)}\right]^2}{\frac{1}{2\pi}e^{-\frac{1}{2}\beta(x,y)}}dx\,dy$$

unde:

$$\alpha(x,y)=\frac{1}{1-r^2}\left(x^2-2rxy+y^2\right)\qquad\beta(x,y)=x^2+y^2.$$

Un calcul direct ne conduce la valoarea:

Scanned with OKEN Scanner

$$\varphi^2 = \frac{r^2}{1-r^2}$$

deci pentru distribuția considerată $C = r$ (§ 154).

221. *Corelația rangurilor.* În cazul distribuțiilor continue, rangul este mai greu de definit în mod explicit. Vom recurge la valorile funcției frecvențelor cumulate. Pentru distribuții normale și pentru frecvențe absolute, vom avea:

$$N_x = \frac{N}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{x^2}{2}} dx . \tag{187}$$

Rangul teoretic al valorii x va fi egal cu ν (întreg) dacă:

$$\nu - 1 < N_x \leq \nu \tag{188}$$

Să calculăm momentele succesive ale funcțiilor N_x și N_y, presupunând că suntem în cazul distribuției normale (186). Avem:

$$\overline{N_x} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} N_x e^{-\frac{x^2}{2}} dx = \frac{1}{N} \int_{-\infty}^{+\infty} N_x dN_x = \frac{N}{2}$$

$$\overline{N_x^2} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} N_x^2 e^{-\frac{x^2}{2}} dx = \frac{1}{N} \int_{-\infty}^{+\infty} N_x^2 dN_x = \frac{N^2}{3} .$$

Prin urmare, abaterea tip a frecvențelor cumulate va fi:

$$\sigma_N^2 = \frac{N^2}{12} \tag{189}$$

iar a rangurilor:

$$\sigma_\nu^2 = \frac{N^2 - 1}{12} \ ^1 \tag{190}$$

termenul $\frac{1}{12}$ reprezentând corecția Sheppard pentru gruparea în intervale egale cu 1 (188).

Pentru calculul momentului produs:

[1] *Ambele aceste formule sunt evident valabile și în cazul general.*

$$\overline{N_x N_y} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} N_x N_y f(x,y)dx\,dy \qquad (191)$$

vom observa că:

$$\frac{df}{dr} = \frac{d^2 f}{dx\,dy}.$$

Derivând în raport cu r în (191), integrând prin părți și observând că părțile dintre limite sunt nule, obținem:

$$\frac{d\overline{N_x N_y}}{dr} = \iint \frac{dN_x}{dx}\frac{dN_y}{dy} f(x,y)dx\,dy =$$

$$= \frac{N^2}{4\pi^2\sqrt{1-r^2}} \iint e^{-\frac{1}{2(1-r^2)}\left[\left(2-r^2\right)x^2 - 2rxy + \left(2-r^2\right)y^2\right]} dx\,dy =$$

$$= \frac{N^2}{2\pi\sqrt{4-r^2}}$$

deci:

$$\overline{N_x N_y} = \frac{N^2}{2\pi}\arcsin\frac{r}{2} + c.$$

Cum în cazul când $r = 0$, $\overline{N_x N_y} = \overline{N_x}\cdot\overline{N_y}$, deducem $c = \frac{N^2}{4}$ și, prin urmare, coeficientul de corelație al frecvențelor cumulate va fi (142):

$$\rho = \frac{6}{\pi}\arcsin\frac{r}{2}.$$

Ținând seama de relația evidentă:

$$\frac{S(x-y)^2}{N} = \mu_{20} + \mu_{02} - 2\mu_{11} \qquad (192)$$

avem, în cazul frecvențelor cumulate:

$$\frac{S(N_x - N_y)^2}{N} = \frac{N^2}{12} + \frac{N^2}{12} - \frac{N^2}{6}\rho$$

deci:

$$\rho = 1 - \frac{6S(N_x - N_y)^2}{N^3}.$$

Pentru ranguri, va trebui să introducem în (192) abaterea tip a acestora (190):

$$\frac{S(\nu_1 - \nu_2)^2}{N} = 2\left(\frac{N^2-1}{12}\right) - \frac{N^2-1}{6}\rho$$

de unde (140):

$$\rho = 1 - \frac{6S(\nu_1 - \nu_2)^2}{N(N^2-1)}.$$

222. *Corelație și regresie parțială*. Se arată cu ajutorul teoriei formelor pătratice, că funcției caracteristice:

$$\varphi(t_1, t_2, \ldots, t_s) = e^{\frac{1}{2}K(t_1, t_2, \ldots t_s)}$$

în care:

$$K(t_1, t_2, \ldots, t_s) = \sum \sigma_i^2 t_i^2 + \sum r_{jk}\sigma_j\sigma_k t_j t_k$$

îi corespunde distribuția:

$$f(x_1, x_2, \ldots, x_s) = \frac{1}{(2\pi)^{\frac{s}{2}}\Delta^{\frac{1}{2}}} e^{-\frac{1}{2}H(x_1, x_2, \ldots x_s)} \qquad (193)$$

unde $H = 1$ și $K = 1$ sunt respectiv ecuațiile punctuale și tangențiale ale aceluiași elipsoid, iar Δ - discriminantul formei K.[1] În acest caz:

$$H = -\frac{1}{\Delta}\begin{vmatrix} & & x_1 \\ & \Delta & x_2 \\ & & \vdots \\ & & x_s \\ x_1 & x_2 \ldots x_s & 0 \end{vmatrix}$$

[1] *G. Darmois, **op. cit.**, p. 202.*

Relaţia (193) defineşte distribuţia normală de s variabile.
Δ_{ij} fiind minorul de ordinul 2 al lui Δ, avem:

$$H = \sum x_i^2 \frac{\Delta_{ii}}{\Delta} + \sum x_j x_k \frac{\Delta_{jk}}{\Delta} =$$

$$= \frac{\Delta_{11}}{\Delta}\left(x_1 + \sum_2^s x_i \frac{\Delta_{1i}}{\Delta_{11}}\right) + H(0, x_2, \ldots, x_s) - \frac{1}{\Delta\Delta_{11}}\left(\sum_2^s x_i \Delta_{1i}\right)^2$$

şi, prin urmare, distribuţia legată de x_2, x_3, ..., x_s este normală; media şi fluctuaţia acesteia vor fi respectiv:

$$x_{1\cdot 23\ldots s} = -\sum_2^s x_i \frac{\Delta_{1i}}{\Delta_{11}} \tag{194}$$

$$\sigma^2_{1\cdot 23\ldots s} = \frac{\Delta}{\Delta_{11}}. \tag{195}$$

Relaţia (194) este ecuaţia spaţiului liniar de regresie a lui x_1 în x_2, x_3, ..., x_s, coeficienţii de regresie fiind:

$$b_{1i\cdot 23\ldots s} = -\frac{\Delta_{1i}}{\Delta_{11}}. \tag{196}$$

Punând acum:

$$R = \begin{vmatrix} 1 & r_{12} & r_{13}\ldots & r_{1s} \\ r_{21} & 1 & r_{23}\ldots & r_{2s} \\ \ldots & \ldots & \ldots\ldots & \ldots \\ r_{s1} & r_{s2} & r_{s3}\ldots & r_{ss} \end{vmatrix}$$

relaţiile (195) şi (196) devin:

$$\sigma^2_{1\cdot 23\ldots s} = \frac{R}{R_{11}} \cdot \sigma_s^2$$

$$b_{1i\cdot 23\ldots s} = -\frac{R_{1i}}{R_{11}} \cdot \frac{\sigma_1}{\sigma_i}. \tag{197}$$

Coeficientul de corelaţie parţială se defineşte ca media geometrică a coeficienţilor de regresie:

Scanned with OKEN Scanner

$$r_{1i} = \sqrt{b_{1i.} \times b_{i1.}} = -\frac{R_{1i}}{\sqrt{R_{11}R_{ii}}}\ ^{1}. \qquad (198)$$

Se constată că ambii coeficienţi de regresie $b_{1i.}$ şi $b_{i1.}$ se anulează în acelaşi timp şi că sunt nuli numai dacă $r_{1i.} = 0$.

Distribuţia liberă de x_1 va fi şi ea o distribuţie normală ce se construieşte pornind de la Δ_{11}.

Aceste rezultate ne permit stabilirea imediată a oricăreia din relaţiile introduse în Capitolul XII. Trebuie totuşi relevat faptul că, astfel pusă problema, ea implică normalitatea distribuţiilor variabilelor considerate. Relaţiile de mai sus pot fi însă obţinute şi pe baza unui alt punct de vedere. Căutând să exprimăm media legată $x_{1\,.\,23\,..s}$ ca o funcţie liniară de $x_2, x_3, \ldots, x_s$, vom determina coeficienţii b astfel ca suma pătratelor abaterilor $\sum\left(x_{1\cdot 23..s} - x_1\right)^2$ să fie un mimimum. Această condiţie determină complet coeficienţii căutaţi, prin sistemul:

$$\sum_{i=2}^{s} d_{1i.}x_i x_k = \sum x_1 x_k \qquad (k = 2,3,\ldots,s).$$

Presupunând că variabilele sunt măsurate în unităţi egale cu abaterile-tip:

$$\frac{\sum x_i^2}{N} = 1 \qquad \frac{\sum x_i x_k}{N} = r_{ik}$$

determinantul sistemului precedent este chiar R_{11}, de unde se deduce că b_{1i} va avea aceeaşi valoare ca mai sus (197). Un alt calcul simplu ne arată că:

$$\frac{\sum\left(x_{1\cdot 23..s} - x_1\right)^2}{N} = \frac{R}{R_{11}} = \sigma^2_{1\cdot 23..s}.$$

Identitatea formală dintre rezultatele valabile în cazul distribuţiilor normale şi cele obţinute pe baza principiului celor mai mici pătrate este astfel stabilită.

223. Coeficientul de satisfacţie P. Să considerăm o urnă în care sunt $s + 1$ feluri de bile în proporţiile $p_1 : p_2 : \ldots : p_{s+1}$. Se fac n extracţii (bila extrasă fiind repusă în urnă); fie n_1', n_2', ..., n'_{s+1} frecvenţele statistice ale diferitelor categorii de bile. Funcţia caracteristică a abaterilor $n_i' - np_i$ va fi:

$$\varphi(t_1,t_2,\ldots,t_s) = e^{-n\sum p_i t_i}\left(p_{s+1} + \sum p_i e^{t_i}\right)^n.$$

[1] *R fiind simetric, avem* $R_{1i} = R_{i1}$.

Adoptând ca unităţi de măsură abaterile tip $\sqrt{np_iq_i}$ şi neglijând termenii în $\frac{1}{\sqrt{n}}$, avem:

$$\log\varphi = \sum \frac{1}{q_i}\cdot\frac{t_i^2}{2} - \frac{1}{2}\left(\sum\sqrt{\frac{p_i}{q_i}}t_i\right)^2$$

şi, revenind la primele unităţi:

$$log\,\varphi = \frac{1}{2}K(t_1,t_2,\ldots,t_s) = \frac{n}{2}\left[\sum p_i t_i^2 - \left(\sum p_i t_i\right)^2\right]$$

Pentru determinarea formei pătratice H corespunzătoare lui K (§ 222), vom utiliza sistemul cunoscut:

$$\frac{1}{2}\frac{\partial K}{\partial t_k} = np_k\left(t_k - \sum p_i t_i\right) = x_k \qquad (k = 1,2,\ldots,s).$$

De aici:

$$\sum_1^s \frac{x_k^2}{np_k} = n\left[\sum p_k t_k^2 - \left(\sum p_k t_k\right)^2 - p_{s+1}\left(\sum p_k t_k\right)^2\right]$$

$$\sum x_k = np_{s+1}\left(\sum p_k t_k\right)$$

deci:

$$K = H = \sum_1^{s+1}\frac{x_k^2}{np_k}$$

unde $x_1 + x_2 + \ldots + x_{s+1} = 0$. Prin urmare, distribuţia corespunzătoare va fi:

$$f(x_1,x_2,\ldots,x_{s+1})dx_1dx_2\ldots dx_s =$$

$$= \frac{1}{\sqrt{s}(2\pi n)^{\frac{s}{2}}\sqrt{p_1p_2\ldots p_{s+1}}}\,e^{-\frac{1}{2}\sum_1^{s+1}\frac{x_k^2}{np_k}}\,dv \qquad (199)$$

dv fiind elementul de volum al spaţiului liniar $\sum_1^{s+1} x_k = 0$, $(dv = \sqrt{s}dx_1dx_2\ldots dx_s)$.

Scanned with OKEN Scanner

Făcând schimbarea de variabile $x_k = \sqrt{np_k}\, y_k$, cum cosinusul unghiului normalei hiperplanului:

$$\sum_{1}^{s+1} \sqrt{p_k}\, y_k = 0 \tag{200}$$

cu axa Oy_{s+1} este $\sqrt{p_{s+1}}$, expresia (199) devine:

$$fdV = \frac{1}{(2\pi)^{\frac{s}{2}}} e^{-\frac{1}{2}\sum y_k^2} dV \tag{201}$$

unde dV este un element de volum al hiperplanului (200). Utilizând coordonatele polare ale acestuia:

$$y_1 = \chi \cos \varphi_1$$
$$y_2 = \chi \sin \varphi_1 \cos \varphi_2$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$y_{s-1} = \chi \sin \varphi_1 \sin \varphi_2 \ldots \sin \varphi_{s-2} \cos \varphi_{s-1}$$

şi înlocuind în (201), avem:

$$fdV = \frac{1}{(2\pi)^{\frac{s}{2}}} e^{-\frac{\chi^2}{2}} \chi^{s-1} \sin^{s-2} \varphi_1 \sin^{s-3} \varphi_2 \ldots \sin \varphi_{s-2} d\varphi_1 d\varphi_2 \ldots d\varphi_{s-1} d\chi.$$

Integrând acum aici în raport cu unghiurile φ, obţinem frecvenţa eşantioanelor a căror imagine geometrică în spaţiul (200) este cuprinsă între sferele de rază χ şi $\chi + d\chi$ Vom observa însă în prealabil că, $B(m, n)$ fiind integrala euleriană de prima speţă:

$$\int_0^{\pi} \sin^k \varphi\, d\varphi = B\left(\frac{k+1}{2}, \frac{1}{2}\right) = \frac{\Gamma\left(\frac{k+1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{k+2}{2}\right)}$$

vom avea:

Scanned with OKEN Scanner

$$\int_0^{\pi} \sin^{s-2}\varphi_1 d\varphi_1 \int_0^{\pi} \sin^{s-3}\varphi_2 d\varphi_2 \ldots \int_0^{\pi} \sin\varphi_{s-2} d\varphi_{s-2} = \frac{\Gamma\left(\frac{1}{2}\right)^{s-2}}{\Gamma\left(\frac{s}{2}\right)}$$

şi, prin urmare, frecvenţa căutată va fi:

$$\frac{e^{-\frac{\chi^2}{2}} \chi^{s-1} d\chi}{2^{\frac{s}{2}-1}\Gamma\left(\frac{s}{2}\right)}$$

iar probabilitatea ca un număr de extracţii să conducă la un eşantion pentru care:

$$\sum \frac{x_i^2}{np_i} = \sum \frac{(n_i' - n_i)^2}{n_i}$$

să fie mai mare decât un χ^2 dat, va fi (§ 87):

$$P = \frac{\int_{\chi}^{\infty} e^{-\frac{\chi^2}{2}} \chi^{s-1} d\chi}{\int_0^{\infty} e^{-\frac{\chi^2}{2}} \chi^{s-1} d\chi}.$$

224. *Distribuţia valorilor statistice*. Orice valoare caracteristică a unui eşantion este de fapt o funcţie de frecvenţele relative statistice, f_i. Să însemnăm cu:

$$\delta f_i = f_i - p_i$$

diferenţele dintre frecvenţele statistice şi cele teoretice, p_i. Funcţia caracteristică a distribuţiei diferenţelor δf_i este:

$$\varphi(t_1 t_2, \ldots, t_s) = e^{-\sum p_i t_i \left(\sum p_i e^{\frac{t_i}{n}}\right)^n}$$

De aici deducem că cea mai mică putere a lui $1/n$ conţinută în orice moment de ordinul k al diferenţelor δf_i este egală cu partea întreagă a lui $(k+1)/2$. Fie $F(f_1, f_2, \ldots, f_s)$ o caracteristică oarecare a eşantionului; aceeaşi caracteristică a colectivităţii teoretice va fi $F(p_1, p_2, \ldots, p_s)$. Abaterea caracteristicii F va fi:

Scanned with OKEN Scanner

$$\delta F(f_1,f_2,\ldots,f_s) = F(f_1,f_2,\ldots,f_s) - F(p_1,p_2,\ldots,p_s) = \sum_i \frac{\partial F}{\partial f_i}\delta f_i +$$

$$+\frac{1}{2!}\sum_i\sum_j \frac{\partial^2 F}{\partial f_i \partial f_j}\delta f_i \delta f_j + \frac{1}{3!}\sum_i\sum_j\sum_k \frac{\partial^3 F}{\partial f_i \partial f_j \partial f_k}\delta f_i \delta f_j \delta f_k + \ldots$$

Ţinând seama de observaţia precedentă, deducem că pentru a obţine o primă aproximaţie este suficient să luăm:

$$\delta_1 F = \sum_i \frac{\partial F}{\partial f_i}\delta f_i . \tag{202}$$

Aproximaţia de gradul imediat superior va necesita întrebuinţarea relaţiei:

$$\delta_2 F = \sum_i \frac{\partial F}{\partial f_i}\delta f_i + \frac{1}{2}\sum_i\sum_k \frac{\partial^2 F}{\partial f_i \partial f_k}\delta f_i \delta f_k .$$

Să considerăm acum o funcţie de mai multe caracteristici $\Phi(F_1,F_2\ldots)$. Vom avea:

$$\delta\Phi = \sum_i \frac{\partial \Phi}{\partial F_i}\delta F_i + \frac{1}{2!}\sum_i\sum_k \frac{\partial^2 \Phi}{\partial F_i \partial F_k}\delta F_i \delta F_k + \ldots$$

de unde, ţinând seama de (202), deducem:

$$\delta_1\Phi = \sum_i \frac{\partial \Phi}{\partial F_i}\delta_1 F_i . \tag{203}$$

Relaţii similare se pot stabili pentru orice grad de aproximaţie. Nu toate acestea sunt tot aşa de simple ca (203)[1].

Din consideraţiile precedente se desprinde noţiunea de *diferenţială statistică*. Aceasta, cel puţin în ceea ce priveşte ordinul 1, prezintă o analogie formală completă cu diferenţialele matematice. Totuşi, cele două noţiuni nu pot fi confundate: *diferenţiala statistică nu pune în evidenţă legătura ei cu infinitul mic principal decât prin intermediul momentelor.*

Să însemnăm prin $\{X_1, X_2, \ldots, X_k\}$ momentul de ordinul k al unei distribuţii de mai multe variabile. Fie m', μ' momentele statistice ale unui eşantion extras dintr-o colectivitate ale cărei momente sunt m şi μ. Vom avea $m'_\alpha = \sum f_i x_i^\alpha$ şi, prin urmare:

---

[1] *N. St. Georgescu,* ***Further Contributions to the Sampling Problem****, Biometrika, vol. XXIV, p. 65.*

$$\delta m'_\alpha = \sum x_i^\alpha \delta f_i$$

De aici deducem:

$$\left\{\delta m'_\alpha \delta m'_\beta\right\} = \sum_i \sum_k x_i^\alpha x_k^\beta \left\{\delta f_i \delta f_k\right\}$$

ori:

$$\left\{\delta f_i \delta f_k\right\} = \begin{cases} \dfrac{p_i q_i}{n} & i = k \\ -\dfrac{p_i q_k}{n} & i \neq k \end{cases}$$

deci:

$$\left\{\delta m'_\alpha \delta m'_\beta\right\} = \frac{1}{n}\left(m_{\alpha+\beta} - m_\alpha m_\beta\right) \tag{204}$$

La fel:

$$\left\{\delta f_i \delta f_j \delta f_k\right\} = \begin{cases} \dfrac{p_i(1-p_i)(1-2p_i)}{n^2} & i = j = k \\ -\dfrac{p_i p_k (1-2p_i)}{n^2} & i = j \neq k \\ \dfrac{2p_i p_j p_k}{n^2} & i \neq j \neq k \end{cases}$$

și deci:

$$\left\{\delta m'_\alpha \delta m'_\beta \delta m'_\gamma\right\} = \frac{1}{n^2}\left(m_{\alpha+\beta+\gamma} - m_{\alpha+\beta} m_\gamma - m_{\beta+\gamma} m_\alpha - m_{\gamma+\alpha} m_\beta + 2 m_\alpha m_\beta m_\gamma\right)$$

Să considerăm acum relația (203) și să presupunem că originea aleasă este chiar media aritmetică a colectivității teoretice (care în general nu este egală cu media eșantionului). Vom avea:

$$\delta_1 \mu'_\alpha = \delta m'_\alpha - \alpha \mu_{\alpha-1} \delta m'_1$$

de unde:

$$\left\{\delta_1 \mu'_\alpha \delta_1 \mu'_\beta\right\} = \left\{\delta m'_\alpha \delta m'_\beta\right\} - \alpha \mu_{\alpha-1}\left\{\delta m'_\beta \delta m'_1\right\} - \beta \mu_{\beta-1}\left\{\delta m'_\alpha \delta m'_1\right\} + \alpha\beta \mu_{\alpha-1}\mu_{\beta-1}\left\{\delta^2 m'_1\right\}$$

Scanned with OKEN Scanner

Ţinând seama de relaţia (204) şi de ipoteza făcută:

$$\left\{\delta_1\mu'_\alpha\delta_1\mu'_\beta\right\}=\frac{1}{n}\left(\mu_{\alpha+\beta}-\mu_\alpha\mu_\beta-\alpha\mu_{\alpha-1}\mu_{\beta+1}-\beta\mu_{\alpha+1}\mu_{\beta-1}+\alpha\beta\mu_{\alpha-1}\mu_{\beta-1}\mu_2\right) \qquad (205)$$

care, când $\alpha=\beta$, conduce la fluctuaţia lui μ'_α:

$$\sigma^2\left(\mu'_\alpha\right)=\frac{1}{n}\left(\mu_{2\alpha}-\mu_\alpha^2-2\alpha\mu_{\alpha-1}\mu_{\alpha+1}+\alpha^2\mu_{\alpha-1}^2\mu_2\right). \qquad (205\text{ bis})$$

deoarece $\left\{\delta_1\mu'_\alpha\right\}=0$. Din (205 bis) deducem prin particularizare parte din relaţiile cuprinse în tabelul 82, § 103.

Utilizând din nou relaţia (203):

$$\delta_1\beta'_2=\frac{\delta'_1\mu'_4}{\mu_2^2}-\frac{2\mu_4\delta_1\mu'_2}{\mu_2^3}$$

de unde, ţinând seama de (205):

$$\sigma^2\left(\beta'_2\right)=\frac{1}{n}\left(\frac{\mu_8}{\mu_2^4}-\frac{4\mu_4\mu_6}{\mu_2^5}+\frac{4\mu_4^3}{\mu_2^6}-\frac{\mu_4^2}{\mu_2^4}-\frac{8\mu_3\mu_5}{\mu_2^5}+\frac{16\mu_3^2\mu_4}{\mu_2^5}+\frac{16\mu_3^2}{\mu_2^3}\right). \qquad (206)$$

În cazul unei distribuţii normale (170):

$$\mu_3=0 \qquad \mu_4=3\mu_2^2 \qquad \mu_6=15\mu_2^3 \qquad \mu_8=105\mu_2^4$$

şi, prin urmare, (206) devine:

$$\sigma^2\left(\beta'_2\right)=\frac{24}{n}.$$

Pe aceeaşi cale se arată că abaterea tip a coeficientului β'_1 este nulă în primă aproximaţie.

225. *Funcţiile tetrachorice*. Coeficientul de corelaţie tetrachorică se obţine prin rezolvarea ecuaţiei (118):

$$\frac{1}{2\pi\sqrt{1-r^2}}\int_h^\infty\int_k^\infty e^{-\frac{1}{2\left(1-r^2\right)}\left(x^2+y^2-2rxy\right)}dx\,dy=\frac{d}{N} \qquad (207)$$

Să punem:

Scanned with OKEN Scanner

$$\delta m'_\alpha = \sum x_i^\alpha \delta f_i$$

De aici deducem:

$$\left\{\delta m'_\alpha \delta m'_\beta\right\} = \sum_l \sum_k x_l^\alpha x_k^\beta \left\{\delta f_l \delta f_k\right\}$$

ori:

$$\left\{\delta f_i \delta f_k\right\} = \begin{cases} \dfrac{p_i q_i}{n} & i = k \\ -\dfrac{p_i q_k}{n} & i \neq k \end{cases}$$

deci:

$$\left\{\delta m'_\alpha \delta m'_\beta\right\} = \frac{1}{n}\left(m_{\alpha+\beta} - m_\alpha m_\beta\right) \tag{204}$$

La fel:

$$\left\{\delta f_i \delta f_j \delta f_k\right\} = \begin{cases} \dfrac{p_i(1-p_i)(1-2p_i)}{n^2} & i = j = k \\ -\dfrac{p_i p_k (1-2p_i)}{n^2} & i = j \neq k \\ \dfrac{2p_i p_j p_k}{n^2} & i \neq j \neq k \end{cases}$$

și deci:

$$\left\{\delta m'_\alpha \delta m'_\beta \delta m'_\gamma\right\} = \frac{1}{n^2}\left(m_{\alpha+\beta+\gamma} - m_{\alpha+\beta} m_\gamma - m_{\beta+\gamma} m_\alpha - m_{\gamma+\alpha} m_\beta + 2 m_\alpha m_\beta m_\gamma\right)$$

Să considerăm acum relația (203) și să presupunem că originea aleasă este chiar media aritmetică a colectivității teoretice (care în general nu este egală cu media eșantionului). Vom avea:

$$\delta_1 \mu'_\alpha = \delta m'_\alpha - \alpha \mu_{\alpha-1} \delta m'_1$$

de unde:

$$\left\{\delta_1 \mu'_\alpha \delta_1 \mu'_\beta\right\} = \left\{\delta m'_\alpha \delta m'_\beta\right\} - \alpha \mu_{\alpha-1}\left\{\delta m'_\beta \delta m'_1\right\} - \beta \mu_{\beta-1}\left\{\delta m'_\alpha \delta m'_1\right\} + \alpha\beta \mu_{\alpha-1}\mu_{\beta-1}\left\{\delta^2 m'_1\right\}$$

Scanned with OKEN Scanner

Ţinând seama de relaţia (204) şi de ipoteza făcută:

$$\left\{\delta_1\mu'_\alpha\delta_1\mu'_\beta\right\}=\frac{1}{n}\left(\mu_{\alpha+\beta}-\mu_\alpha\mu_\beta-\alpha\mu_{\alpha-1}\mu_{\beta+1}-\beta\mu_{\alpha+1}\mu_{\beta-1}+\alpha\beta\mu_{\alpha-1}\mu_{\beta-1}\mu_2\right) \qquad (205)$$

care, când $\alpha=\beta$, conduce la fluctuaţia lui μ'_α:

$$\sigma^2\left(\mu'_\alpha\right)=\frac{1}{n}\left(\mu_{2\alpha}-\mu_\alpha^2-2\alpha\mu_{\alpha-1}\mu_{\alpha+1}+\alpha^2\mu_{\alpha-1}^2\mu_2\right). \qquad (205\text{ bis})$$

deoarece $\left\{\delta_1\mu'_\alpha\right\}=0$. Din (205 bis) deducem prin particularizare parte din relaţiile cuprinse în tabelul 82, § 103.

Utilizând din nou relaţia (203):

$$\delta_1\beta'_2=\frac{\delta'_1\mu'_4}{\mu_2^2}-\frac{2\mu_4\delta_1\mu'_2}{\mu_2^3}$$

de unde, ţinând seama de (205):

$$\sigma^2\left(\beta'_2\right)=\frac{1}{n}\left(\frac{\mu_8}{\mu_2^4}-\frac{4\mu_4\mu_6}{\mu_2^5}+\frac{4\mu_4^3}{\mu_2^6}-\frac{\mu_4^2}{\mu_2^4}-\frac{8\mu_3\mu_5}{\mu_2^5}+\frac{16\mu_3^2\mu_4}{\mu_2^5}+\frac{16\mu_3^2}{\mu_2^3}\right). \qquad (206)$$

În cazul unei distribuţii normale (170):

$$\mu_3=0 \qquad \mu_4=3\mu_2^2 \qquad \mu_6=15\mu_2^3 \qquad \mu_8=105\mu_2^4$$

şi, prin urmare, (206) devine:

$$\sigma^2\left(\beta'_2\right)=\frac{24}{n}.$$

Pe aceeaşi cale se arată că abaterea tip a coeficientului β'_1 este nulă în primă aproximaţie.

225. *Funcţiile tetrachorice*. Coeficientul de corelaţie tetrachorică se obţine prin rezolvarea ecuaţiei (118):

$$\frac{1}{2\pi\sqrt{1-r^2}}\int_h^\infty\int_k^\infty e^{-\frac{1}{2\left(1-r^2\right)}\left(x^2+y^2-2rxy\right)}dx\,dy=\frac{d}{N} \qquad (207)$$

Să punem:

$$F=\frac{1}{\sqrt{1-r^2}}e^{-\frac{1}{2(1-r^2)}(x^2+y^2-2rxy)}=e^{-\frac{1}{2}(x^2+y^2)}\sum_0^\infty u_i\frac{r^i}{k!}.$$

Este uşor de arătat că:

$$u_l=p_i(x)p_i(y) \tag{208}$$

unde:

$$p_k(x)=xp_{k-1}-(k-1)p_{k-2} \qquad p_0=1 \quad p_1=x \tag{209}$$

şi că:

$$\frac{dp_k}{dx}=kp_{k-1}.$$

Prin urmare (§ 155):

$$\frac{1}{2\pi}\int_h^\infty\int_k^\infty F dxdy=\sum_0^\infty \tau_i(h)\tau_i(k)\frac{r^i}{i!}$$

unde am pus:

$$\tau_i(h)=\frac{1}{\sqrt{2\pi}}\int_h^\infty e^{-\frac{x^2}{2}}p_i(x)dx.$$

Din relaţiile (208) şi (209) deducem:

$$p_i=xp_{i-1}-\frac{dp_{i-1}}{dx}$$

şi, prin urmare:

$$\tau_i(h)=\frac{1}{\sqrt{2\pi}}\int_h^\infty xe^{-\frac{x^2}{2}}p_{i-1}(x)dx-\frac{1}{\sqrt{2\pi}}\int_0^\infty e^{-\frac{x^2}{2}}dp_{i-1}(x).$$

Integrând prin părţi ultima integrală, avem, cu notaţiile din § 157:

$$\tau_i(h)=\frac{1}{\sqrt{2\pi}}e^{-\frac{h^2}{2}}p_{i-1}(h)=Hp_{i-1}(h) \qquad (i\geq 1).$$

Scanned with OKEN Scanner

Funcţiile tetrachorice τ_i au fost introduse de prof. K. PEARSON. Ele sunt strâns legate de *polinoamele tetrachorice*, p_i.

Utilizând această relaţie, ecuaţia (207) se poate pune sub forma:

$$\frac{ad - bc}{N^2 HK} = \sum_1^{\infty} p_{i-1}(h) p_{i-1}(k) \frac{r^2}{i!}.$$

Scanned with OKEN Scanner

Table numerice

TABLA I

Valorile funcției $z = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

Partea zecimală

X	0	1	2	3	4	5	6	7	8	9
0,0	39894	39892	39886	39876	39862	39844	39822	39797	39767	39733
0,1	39695	39654	39608	39559	39505	39448	39387	39322	39253	39181
0,2	39104	39024	38940	38853	38762	38667	38568	38466	38361	38251
0,3	38139	38023	37903	37780	37654	37524	37391	37255	37115	36973
0,4	36827	36678	36526	36371	36213	36053	35889	35723	35553	35381
0,5	35207	35029	34849	34667	34482	34294	34105	33912	33718	33521
0,6	33322	33121	32918	32713	32506	32297	32086	31873	31659	31443
0,7	31225	31006	30785	30563	30339	30114	29887	29659	29431	29200
0,8	28969	28737	28504	28269	28034	27798	27562	27324	27086	26848
0,9	26608	26369	26129	25888	25647	25406	25164	24923	24681	24439
1,0	24197	23955	23713	23471	23230	22988	22747	22506	22265	22025
1,1	21785	21546	21307	21069	20831	20594	20357	20121	19886	19652
1,2	19419	19186	18954	18723	18494	18265	18037	17810	17585	17360
1,3	17137	16915	16694	16474	16255	16038	15822	15608	15395	15183
1,4	14973	14764	14556	14350	14146	13943	13742	13542	13343	13147
1,5	12952	12758	12566	12376	12188	12001	11816	11632	11450	11270
1,6	11092	10915	10741	10567	10396	10226	10057	09893	09728	09566
1,7	09405	09246	09089	08933	08779	08628	08478	08329	08283	08038
1,8	07895	07754	07614	07477	07341	07206	07074	06943	06814	06687
1,9	06562	06438	06316	06195	06077	05959	05844	05730	05618	05508
2,0	05399	05292	05186	05082	04980	04879	04780	04682	04586	04491
2,1	04398	04307	04217	04182	04041	03955	03871	03788	03706	03626
2,2	03547	03470	03394	03319	03246	03174	03103	03034	02965	02898
2,3	02833	02768	02705	02643	02582	02522	02463	02406	02349	02294
2,4	02239	02186	02134	02083	02033	01984	01936	01889	01842	01797
2,5	01753	01709	01667	01625	01585	01545	01506	01468	01431	01394
2,6	01358	01323	01289	01256	01223	01191	01160	01130	01100	01071
2,7	01042	01014	00987	00961	00935	00909	00885	00861	00837	00814
2,8	00792	00770	00748	00727	00707	00687	00668	00649	00631	00613
2,9	00595	00578	00562	00545	00530	00514	00499	00484	00471	00457
3,0	00443	00430	00417	00405	00393	00381	00370	00358	00348	00337
3,1	00327	00317	00307	00298	00288	00279	00271	00262	00254	00246
3,2	00238	00231	00224	00216	00210	00203	00196	00190	00184	00178
3,3	00172	00167	00161	00156	00151	00146	00141	00136	00132	00128
3,4	00123	00119	00115	00111	00107	00104	00100	00097	00094	00090
3,5	00087	00084	00081	00078	00076	00073	00071	00068	00066	00063
3,6	00061	00059	00056	00055	00053	00051	00049	00047	00046	00044
3,7	00042	00041	00039	00038	00037	00035	00034	00033	00031	00030
3,8	00029	00028	00027	00026	00025	00024	00023	00022	00021	00021
3,9	00020	00019	00018	00018	00017	00016	00016	00015	00014	00014

Scanned with OKEN Scanner

TABLA II

Valorile funcției $S_1(X) = \frac{1}{\sqrt{2\pi}} \int_0^X e^{-\frac{X^2}{2}} dX$

Partea zecimală

X	0	1	2	3	4	5	6	7	8	9
0,0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0,1	03983	04379	04776	05172	05567	05962	06356	06749	07142	07535
0,2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0,3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0,4	15542	15910	16276	16640	17003	17364	17724	18082	18439	18793
0,5	19146	19497	19847	20194	20540	20884	21226	21566	21904	22240
0,6	22575	22907	23237	23566	23891	24215	24537	24857	25175	25490
0,7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0,8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0,9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1,0	34134	34375	34614	34850	35083	35314	35543	35769	35993	36214
1,1	36433	36650	36864	37076	37286	37493	37698	37899	38100	38298
1,2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1,3	40320	40490	40658	40824	40988	41149	41309	41466	41621	41774
1,4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1,5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1,6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1,7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1,8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1,9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2,0	47725	47779	47831	47882	47932	47982	48030	48077	48124	48169
2,1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2,3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2,4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2,7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2,8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2,9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3,0	49865	49869	49874	49878	49882	49886	49889	49893	49897	49900
3,1	49903	49906	49910	49913	49916	49918	49921	49924	49926	49929
3,2	49931	49934	49936	49938	49940	49942	49944	49946	49948	49950
3,3	49952	49953	49955	49957	49958	49960	49961	49962	49964	49965
3,4	49966	49968	49969	49970	49971	49972	49973	49974	49975	49976
3,5	49977	49978	49978	49979	49980	49981	49981	49982	49983	49983
3,6	49984	49985	49985	49986	49986	49987	49987	49988	49988	49989
3,7	49989	49990	49990	49990	49991	49991	49992	49992	49992	49992
3,8	49993	49993	49993	49994	49994	49994	49994	49995	49995	49995
3,9	49995	49995	49996	49996	49996	49996	49996	49996	49997	49997

Scanned with OKEN Scanner

TABLA III

$\pi = 3{,}14159$	$\log \pi = 0{,}49715$
$\frac{1}{2\pi} = 0{,}15915$	$\log \frac{1}{\sqrt{2\pi}} = \bar{1}{,}60091$
$\frac{1}{\sqrt{2\pi}} = 0{,}39894$	$\log e = 0{,}43429$
$e = 2{,}71828$	$\sqrt{2} = 1{,}41421$
$\frac{1}{e} = 0{,}36788$	$\frac{1}{\sqrt{2}} = 0{,}70711$

$$\sqrt{\frac{2}{\pi}} = 0{,}79788$$

Scanned with OKEN Scanner

EDITURA Expert
București, România
Licența Ministerului Culturii nr. 1442/1992
Consilier editorial: *Valeriu IOAN-FRANC*
Redactor: *Ecaterina IONESCU, Paula NEACȘU*
Procesare text: *Mihaela BURUGĂ, Carmen DIȚULESCU*
Tehnoredactor: *Nicolae LOGIN*
Depozit legal trim. II 1998
ISBN 973-9282-37-7
973-9282-38-5

Scanned with OKEN Scanner

În colecția
„Biblioteca
Băncii Naționale"
Nicholas GEORGESCU-ROEGEN
OPERE COMPLETE

I **Omul și opera** (apărut 1996)

II **Economia României**
(apărut 1997)

III **Statistică și matematică**
* **Metoda statistică**
(apărut 1998)
** **Statistică și matematică**
(apărut 1998)

IV **Economia analitică**

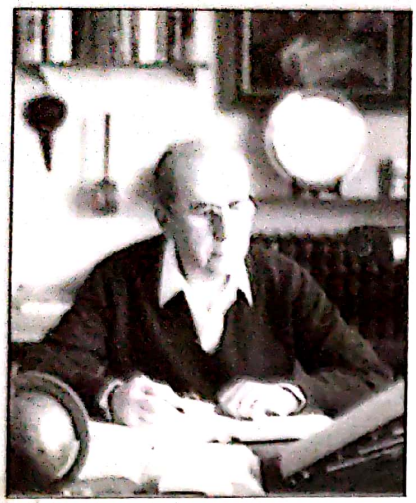
V **Legea entropiei și procesul economic** (apărut 1996)

VI **Energia, resursele naturale și teoria economică**

VII **Epistemologia roegeniană**

Scanned with OKEN Scanner

ISBN 973-9282-37-7; 973-9282-38-5 Apărut 1998



Într-un anume sens, se poate spune că sunt atâtea statistici câți statisticieni sunt.

Sub hainele variate ale feluritelor aspecte ale statisticii sunt un număr de elemente comune, care constituie schema științei statistice. Metodă, în primul rând, adaptată însă în așa măsură universului, încât ambiționează să-i pătrundă cele mai ascunse și mai complexe aspecte și să realizeze cunoașterea lui integrală. Cu valori tipice, cu frecvențe, cu fluctuații, cu corelații și cu indici, cu câteva scheme tipice, statistica speră să dea imaginea lumii, făcând să renască, pentru omul de știință, iluziile pe care le-a trăit mecanicismul veacului trecut. Cu mai multă înțelepciune însă, făcând loc și acelor necesare limite ale cunoașterii științifice în mic, spre lumea microscopică, în care principiul nedeterminării lui Heisenberg sau însăși teoria cuantelor stabilesc marginile posibile ale măsurătorilor, în mare, spre lumea numerelor mari, care constituie prin ea însăși obiect nou de cunoștință diferit de elementele constitutive și aproape esențial străin acestora, deci eventual însăși comprehensiunii lor.

Pornind de la aceste postulate, lucrarea de față prezintă materialul statisticii într-o înlănțuire sistematică și organic gândită.

Ea cuprinde noțiunile statistice fundamentale, metodele statistice cele mai obișnuite, atât relativ la descrierea fenomenelor statistice și analiza lor, cât și la cercetarea relațiilor statistice dintre ele, cu referință specială la unele fenomene economice și la acelea care constituie biometrica, fenomene care formează nu numai un obiect de preferință a autorului, dar și un model al oricărei statistici.

Cartea de față reprezintă în esență cursul de statistică matematică pe care autorul îl face, din însărcinarea mea, la Școala de Statistică.

Studiile teoretice, aprofundate și variate în același timp, cu care dl. N. Georgescu-Roegen, statistician de la Institutul de Statistică al Universității din Paris, a trecut prin University College de la Londra, lucrând cu Pearson, l-au indicat de la sine pentru o atare însărcinare, ca și pentru aceea reprezentată de această lucrare.

OCTAV ONICESCU
Prefață la prima ediție, 1933

colecția BIBLIOTECA BĂNCII NAȚIONALE

13

Nicholas
GEORGESCU-ROEGEN

Metoda statistică

Editura Expert

III